AFCP

ATALA

JEP-TALN-RECITAL 2016

Journées d'Études sur la Parole
Traitement Automatique des Langues Naturelles
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues

PARIS Inalco du 4 au 8 juillet 2016

Organisé par les laboratoires franciliens

https://jep-taln2016.limsi.fr

Conférenciers invités:

Christian Chiarcos (Goethe-Universität, Frankfurt.)
Mark Liberman (University of Pennsylvania, Philadelphia)

Coordinateurs comités d'organisation

Nicolas Audibert et Sophie Rosset (JEP)
Laurence Danlos & Thierry Hamon (TALN)
Damien Nouvel & Ilaine Wang (RECITAL)
Philippe Boula de Mareuil, Sarra El Ayari & Cyril Grouin (Ateliers)

Présentation de l'atelier ELTAL : Enseignement des langues et TAL

Ivan Smilauer[1] Jovan Kostov[1]
(1) E.A. 4514 – PLIDAM, 2, rue de Lille, 75007 Paris, France
ivan.smilauer@inalco.fr, jovan.kostov@gmail.com

RÉSUMÉ

ELTAL est un atelier organisé au sein de la conférence JEP-TALN-RECITAL 2016 et regroupe des contributions qui étudient les méthodes et les applications en TAL dans le domaine de l'enseignement des langues.

ABSTRACT

Language Teaching and NLP.
ELTAL is a workshop organized within the JEP-TALN-RECITAL 2016 conference. This workshop brings together papers investigating metdods and applications in NLP applied to language teaching.

MOTS-CLÉS : TAL, didactique, enseignement, langues, applications, modélisations.
KEYWORDS: NLP, didactics, teaching, languages, applications, modelizations.

Aujourd'hui, l'usage des TICE devient incontournable dans le domaine de l'enseignement et de l'apprentissage des langues. Les applications de partage des connaissances, des nouvelles formes de e-learning telles que MOOC ou SPOC, ainsi que la banalisation des appareils mobiles connectés contribuent à la prolifération des outils numériques destinés aux apprenants de langues (ALAO). Ces outils sont généralement conçus dans une optique de simple adaptation des contenus didactiques traditionnels au support numérique, sans creuser les possibilités qui s'offrent grâce au traitement automatique des langues, comme dans toute application ayant trait au langage naturel. Ce sont justement ces possibilités que nous mettons en valeur dans cet atelier.

Une décennie après l'atelier « TAL et ALAO » organisé à Leuven dans le cadre de TALN 2006, il nous a semblé opportun de proposer cette rencontre qui a pour but de faire un état de l'art des interactions entre didactique des langues et TAL et d'étudier de nouvelles perspectives de la modélisation linguistique. L'objectif sous-jacent est de présenter des outils et des ressources pour l'enseignement des langues au sens large (étrangères ou maternelles) et sans restriction à une aire linguistique ou géographique. Nous avons voulu en particulier tirer profit du fait que la conférence TALN ait pour cadre l'INALCO, qui propose l'enseignement d'une centaine de langues du monde entier et qui représente un environnement privilégié pour créer un moment de partage entre spécialistes du TAL et didacticiens et mettre en évidence les points de rencontre entre ces deux disciplines.

L'atelier est articulé en deux axes en fonction de la nature des données langagières traitées. Le premier axe porte sur l'enrichissement des ressources didactiques et est orienté sur les traitements spécifiques opérés sur la langue cible dans sa forme standardisée. Le second axe porte sur l'analyse des productions langagières des apprenants. Cette analyse prend en compte le caractère non-standard des productions, étudié notamment dans le domaine de l'acquisition d'une langue étrangère pour repérer les zones de difficulté et identifier les besoins de l'apprenant.

Nous avons appelé à soumettre des interventions qui présentent des travaux théoriques et/ou des applications concrètes s'articulant autour des thématiques suivantes :

Axe 1 :

- sélection automatique de ressources textuelles en fonction du genre, de la thématique, de la lisibilité, etc.
- présentation d'informations métalinguistiques produites à l'aide des outils d'analyse automatique
- génération automatique d'exercices à partir d'un corpus
- application didactique des outils de génération (outils de flexion, de dérivation, etc.)
- création et intégration des lexiques et dictionnaires à visée pédagogique
- intégration didactique de la traduction automatique
- aides à la lecture
- aides à l'écriture
- jeux avec un but (gwap)
- synthèse vocale

Axe 2 :

- correction automatique et diagnostic des erreurs
- feedback didactique
- modélisation de l'apprenant par l'étude de ses productions
- constitution et annotation/analyse des corpus d'apprenants
- reconnaissance de la parole.

L'atelier comprend douze interventions avec des thématiques réparties entre les deux axes susmentionnés.

Parmi les contributions acceptées, nous constatons un équilibre entre les articles présentant des problématiques relatives à la classification des apprenants et à l'analyse de leur productions, ceux qui traitent de l'apprentissage d'un domaine particulier (phonologie, phonétique, prosodie, syntaxe) et des outils de génération d'exercices.

Ces contributions reflètent des travaux de recherches menés dans un cadre francophone, mais traitent des aspects et des langues très divers en allant de l'anglais jusqu'au thaï ou l'hindi ou, encore, la langue de signes française. Une part considérable des contributions concerne l'apprentissage des langues en milieu scolaire et universitaire. L'atelier comprend également une conférence invitée (Georges Antoniadis, LIDILEM, Université de Grenoble-Alpes) qui tâche à présenter un état de l'art de l'articulation entre le traitement automatique et la didactique des langues.

Nous espérons que cet atelier contribuera à démarrer une nouvelle dynamique dans le domaine du TAL et de l'enseignement / apprentissage des langues.

Table des matières

Classification d'apprenants francophones de l'anglais sur la base des métriques de complexité lexicale et syntaxique
Nicolas Ballier, Thomas Gaillat 1

De l'exemple construit à l'exemple attesté : un système de requêtes syntaxiques pour non-spécialistes
Ilaine Wang, Sylvain Kahane, Isabelle Tellier 15

D'un corpus à l'identification automatique d'erreurs d'apprenants
Marie-Paule Jacques 22

Du TAL dans les écrits scolaires : premières approches
Claire Wolfarth, Claude Ponton, Catherine Brissaud 30

Élaboration semi-automatique d'une ressource de patrons verbaux
Sylvain Hatier, Rui Yan 38

Exploitation d'une base lexicale dans le cadre de la conception de l'ENPA Innovalangues
Mathieu Mangeot, Valérie Bellynck, Emmanuelle Eggers, Mathieu Loiseau, Yoann Goudin 48

Génération d'exercices d'apprentissage de langue de spécialité par l'exploration du corpus
François-C. Rey, Izabella Thomas, Iana Atanassova 65

Origines des erreurs en Traduction Spécialisée : différentiation textométrique grâce aux corpus de textes cibles annotés
Natalie Kübler, Maria Zimina, Serge Fleury 77

Patrons de coarticulation des voyelles françaises quantiques /i, a, u/ prononcées par des apprenants tchécophones. Illustration du logiciel VisuVo.
Nikola Maurová Paillereau 89

Pratique de la lecture en thaï et hindi en L2 : classification automatique de textes par progression lexicale
Jennifer Lewis-Wong, Satenik Mkhitaryan 103

Un logiciel pour l'enseignement de la prosodie
Philippe Martin 116

Vers une indexation adaptée des ressources pédagogiques sur une plateforme dédiée à l'enseignement de la Langue des Signes Française
Lucie Metz, Virginie Zampa, Saskia Mugnier 124

Classification d'apprenants francophones de l'anglais sur la base des métriques de complexité lexicale et syntaxique

Nicolas Ballier[1] Thomas Gaillat[1, 2]
(1) CLILLAC-ARP (EA3967), rue Thomas Mann, 75013 PARIS, FRANCE
(2) SCELVA, Campus Beaulieu 35042 RENNES cedex, FRANCE
nicolas.ballier@univ-paris-diderot.fr, thomas.gaillat@univ-rennes1.fr

RESUME

Cette contribution examine les monologues en oral spontané du corpus ANGLISH (Tortel 2009). Les productions orales de vingt locuteurs natifs sont comparées aux monologues produits par les quarante locuteurs francophones de niveau intermédiaire et avancé. Les métriques de complexité syntaxique et lexicale implémentées dans des analyseurs (Lu 2014) sont utilisées pour essayer de classer les locuteurs. Enfin, à partir des courbes de croissance du vocabulaire et des modèles LNRE (Baayen 2008), on cherche à évaluer la pertinence de ces métriques de l'écrit pour la classification des locuteurs en fonction de leur production orale.

ABSTRACT

Classifying French learners of English with written-based lexical and complexity metrics.
This paper assesses spontaneous oral monologues in the ANGLISH corpus (Tortel 2009). Twenty oral transcriptions of NS English are compared with forty French-L1 transcriptions of NNS English of intermediate and advanced levels. Syntactic and complexity metrics (Lu 2014) and Vocabulary Growth Curves (Evert & Baroni 2008, Baayen 2008) are used to classify speakers. We analyse how significant these written-based metrics are in the classification of speakers for their oral production.

MOTS-CLES : complexité lexicale, complexité syntaxique, métriques d'apprenants, modèles LNRE

KEYWORDS: syntactic complexity, lexical complexity, vocabulary growth curves, LNRE models.

1 Introduction

Les corpus d'apprenants oraux sont encore rares (Ballier & Martin 2015), alors que bien des corpus d'apprenants écrits ont fait l'objet d'analyses de traitements automatiques (Cf., entre autres, Díaz et al. 2013; Tono 2013). Pour bien des corpus d'apprenants (ANGLISH, NOCE, LeAP), c'est le niveau d'étude à l'université des apprenants qui sert de référence à l'évaluation de leur niveau initial. Il y a donc une sorte de présomption d'homogénéité de la promotion d'étudiants, l'appartenance à un niveau universitaire devant garantir le niveau des sujets enregistrés et étudiés. Dans le cas du corpus oral ANGLISH (Tortel 2009), le choix des sujets enregistrés pour le corpus s'est de plus un peu fait sur la base des situations professionnelles, puisque les locuteurs de niveau intermédiaire (FR1) ont été en partie recrutés sur le campus de l'université (et non sur la base de tests spécifiques, mais aucun n'avait étudié l'anglais en études supérieures), les étudiants

spécialistes d'anglais en troisième année à l'université ont constitué le groupe de locuteurs francophones avancés en anglais (FR2). Le corpus a échantillonné trois populations de locuteurs (GB : natifs anglophones, FR1 et FR2) et permet des comparaisons éclairantes sur des tâches identiques. Pour l'analyse du texte lu (enjeu de la thèse), l'analyse révèle une certaine forme d'homogénéité, au moins en termes de variance, au sein des trois groupes ; le pari global de niveaux d'étudiants conçus à partir de promotion n'est donc pas complètement invalidé. Pour autant, les compétences en anglais des sujets moins avancés (FR1) sont sans doute extrêmement variables, qu'il s'agisse d'un expert lisant régulièrement en anglais dans son domaine ou de sujets ayant conservé une certaine proximité professionnelle en raison de leur activité professionnelle avec les textes écrits en anglais. Dans quelle mesure peut-on utiliser les métriques de complexité lexicale et syntaxique comme des procédures *a posteriori* établissant le niveau des apprenants ?

La conversion en niveaux de référence du CECRL est encore programmatique (mais voir Ovtcharov et al. 2006, ainsi que Leclercq et al. 2014) et cela constitue en quelque sorte l'horizon de ce travail, qui s'inscrit dans la recherche des traits critériés (Hawkins & Filipovic 2012). Sur la base des performances lexicales et syntaxiques, on cherche à distinguer des productions des natifs et des non-natifs, et surtout des stades d'interlangue chez les non-natifs (Gaillat 2013), voire des profils d'apprenants (Chitez 2014). La deuxième partie présente le corpus analysé et la méthode suivie. La troisième section expose les résultats, que la quatrième section discute et met en perspective.

2 Méthode

Cette expérimentation va appliquer à des données orales des métriques mises au point pour l'écrit. Nous présentons successivement les données étudiées et les différentes méthodes d'analyse suivies, par ordre décroissant d'intervention sur les données.

2.1 Description du corpus

Le corpus ANGLISH compte 60 locuteurs : 20 anglophones natifs, 20 francophones natifs ayant arrêté l'anglais au bac (FR01) et 20 étudiants en troisième année de Licence d'anglais à l'Université de Provence (FR02). Une partie en texte lu a fait l'objet d'une thèse qui a cherché à classer les locuteurs sur la base des différences de rythme (Tortel 2009). Le corpus a été transcrit dans les annexes de cette thèse, en suivant des conventions de transcription assez standard pour un corpus d'oral (transcription de pauses pleines, absence de majuscule). Etablies par une spécialiste de l'oral, ces données ne sont pas directement interprétables par les logiciels de l'écrit (nombreux problèmes de casse, de répétitions pour le *parsing* des phrases, étapes indispensables pour les analyses automatiques de la complexité syntaxique). Nous avons dû modifier le format des données pour l'analyseur de complexité : nous avons ponctué le texte et supprimé les répétitions. L'accord inter-annotateur entre les deux experts a été de 85%, les différences portant essentiellement sur les virgules après les adverbiales. Pour l'analyseur de complexité lexicale, il y a une procédure de lemmatisation ; pour les courbes de croissance du vocabulaire, nous avons utilisé les données brutes.

Pour une même consigne ("vous allez parler librement en anglais pendant deux minutes sur le sujet de votre choix. Vous entendrez un signal de "top départ" et je vous ferai signe lorsque les deux minutes seront atteintes. Je vous laisse réfléchir pendant une minute afin de choisir votre thème, si vous n'avez pas d'idées je vous propose de raconter vos dernières vacances"), les trois groupes de locuteurs s'exécutent différemment : 5 027 tokens pour les FR1, 5 683 pour les FR2 et 8 293 pour les natifs. On a prélevé un échantillon comparable des 5 027 tokens dans les trois corpus. Nous avons suivi des exemples de

Baayen 2008 et en quelque sorte «redécoupé » les corpus à des dimensions identiques, une randomisation des occurrences retenues pourrait donner des résultats différents.

2.2 Les métriques de complexité syntaxique

Comme expliqué dans (Lu 2010, Lu 2016) et dans la documentation, l'analyseur de complexité syntaxique L2SCA repose sur le parser de Stanford (Klein & Manning, 2003), génère les fréquences des principales unités textuelles et syntaxiques et calcule les indices de complexité syntaxique en L2 proposés dans la littérature et notamment ceux compilés dans (Wolfe-Quintero et al. 1998). Il produit donc pour chaque fichier texte en entrée une série de 23 traits :

- neuf mesures syntaxiques : le nombre de mots (W), de phrases (S), de syntagmes verbaux (VP), de propositions (C), de T-units (T), de propositions subordonnées (DC), de T-units complexes (CT), de syntagmes coordonnés (CP), et de SN complexes (CN). L'unité centrale à ces mesures est la T-unit, qui est ainsi définie : ''one main clause plus any subordinate clause or nonclausal structure that is attached to or embedded in it'' Hunt (1970:4).

- quatorze indices de complexité syntaxique : la longueur moyenne de la phrase (MLS), la longueur moyenne de la T-unit (MLT), la longueur moyenne de la proposition (MLC), le nombre de propositions par phrase (C/S), le nombre de syntagme verbal par T-unit (VP/T), le nombre de proposition par T-unit (C/T), le nombre de propositions subordonnées par propositions (DC/C), le nombre de propositions dépendantes par T-unit (DC/T), le nombre de T-units par phrase (T/S), le ratio nombre de T-units complexes /nombre de T-Unit (CT/T), le nombre de syntagmes coordonnés par T-unit (CP/T), le nombre de syntagmes coordonnés par proposition (CP/C), le nombre de groupes nominaux complexes par T-unit (CN/T), et le nombre de groupes nominaux complexes par proposition (CP/C). On voit qu'il s'agit d'un ensemble de ratios calculés par rapport à la proposition ou à la T-unit. Cette *T(erminable)- unit* met en avant un concept de clôture du vouloir-dire. Il existe aussi une caractérisation en un sens encore plus lâche, qui comptabilise de la sorte toute proposition indépendante et ses modificateurs, toute proposition qui ne serait pas une indépendante mais serait ponctuée comme une phrase ou bien même des formes impératives (Schneider & Connor, 1990, cité dans Paris 2015 :203-4).

2.3 Traits de la complexité lexicale

Nous avons utilisé l'analyseur de complexité lexicale développé par Xiaofei Lu (Lu 2012). Ce *Lexical Complexity Analyzer* (LCA) est décrit dans plusieurs travaux de ses travaux (Lu 2013, Lu 2014) et à été mis au point pour l'anglais. Il suppose un choix de variété de référence, britannique ou américaine, les fréquences lexicales de référence ayant été calculées (respectivement) à partir du British National Corpus ou de l'American National Corpus.

L'analyseur LCA suppose en entrée un format de données lemmatisées (au format TreeTagger) et produit 24 mesures. Nous reportons ici par commodité les principaux indices et renvoyons à la lecture de (Lu 2012) pour une discussion critique de ces différents indices. Les mesures portent sur la densité lexicale, la variation lexicale et l'élaboration lexicale (*sophistication*). Les mesures de densité lexicale portent sur les types, les tokens et sur le ratio de certaines catégories. La variation lexicale est essentiellement comptabilisée à partir du *type-to-token ratio* et de ses variantes. Les mesures d'élaboration lexicale reposent sur des proportions des unités lexicales plus élaborées.

Métriques	Signification
Sentence	Le nombre de segments séparés par un point dans chaque transcription
Wordtypes, swordtypes	Trois indicateurs dénombrant le nombre de mots en fonction de leur catégorie grammaticale
Lextype, slextypes	Indicateurs du nombre de lemmes lexicaux
Wordtokens, swordtokens	Indicateurs du nombre de tokens liés aux mots différents
lextokens, slextokens	Indicateurs du nombre de tokens lexicaux liés aux lemmes différents
Ld	Indice de densité lexicale (= nombre de mots lexicaux et non grammaticaux/nombre de tokens)
ls1, ls2	Proportions de lexèmes parmi les 2000 mots les plus fréquents (Laufer & Nation 1994)
vs1, vs2, cvs1	Ratio de verbes ne figurant pas parmi les 20 ou 200 verbes les plus fréquents en français / nombre de verbes utilisés (et ses variantes, cf. Wolfe-Quintero et al. 1998)
ndw, ndwz, ndwerz, ndwesz	Indices fondés sur le nombre de mots différents pris dans des échantillons de 50 items.
ttr, msttr, cttr, rttr, logttr, Uber	Indices fondés sur le ratio entre le nombre de mots différents et le nombre total de mots (*type-to-token* ratio et ses avatars normalisés ou transformés :TTR, MSTTR, CTTR, RTTR, et LogTTR). L'index Uber est le Log2(types)/log(types/tokens).
lv, vv1, svv1, cvv1, vv2	Indices de variation concernant le lexique et les formes verbales
nv, adjv, advv, modv	Indices de variation concernant les noms, adjectifs, adverbes et modificateurs (englobant adjectifs et adverbes)

TABLE 1 : Synthèse des principales métriques implémentées dans LCA (Lu 2012, Lu2014).

La réflexion sur la complexité lexicale est dans ce cadre indissociable de la lemmatisation, contrairement aux techniques suivantes.

2.4 L'étude des courbes de croissance du vocabulaire

Bien que pratiquées sur des échelles de données singulièrement différentes (la partie étudiée ici du corpus ANGLISH compte 19 003 tokens, à comparer avec le million du corpus Brown), l'intérêt de la problématique demeure : dans quelle mesure les données d'apprenants, sujettes à l'appauvrissement du lexique au fur et à mesure de leur expression (Bentz et al. 2013) sont-elles justiciables de la loi de Zipf (Zipf 1949) ? L'examen individuel des courbes d'accroissement du vocabulaire de chaque apprenant insisterait sur les spécificités. L'objectif général est de modéliser, à partir de spectres de fréquences, les courbes de l'accroissement du vocabulaire (nombre de types et d'hapax) en fonction de la taille de l'échantillon (nombre de *tokens*, voir Baayen 2008 et, pour une description rapide en français, Turenne 2016). Le spectre fréquentiel est ainsi défini " A frequency spectrum summarizes a frequency distribution in terms of number of types (Vm) per frequency class (m), i.e., it reports how many distinct types occur once, how many types occur twice, and so on. " (Baroni & Evert 2014:2). On a donc établi pour les trois groupes de locuteurs la courbe de croissance du vocabulaire, qui est en partie fondée sur une mesure des hapax (Evert & Baroni 2004, Baroni 2008).

Ensuite, nous avons utilisé la fonction de comparaison de richesse lexicale "compare.richness.fnc" implémentée dans le package {languageR} (Baayen 2008), qui compare les documents deux à deux à partir de leurs spectres fréquentiels. Dans leurs analyses, (Evert & Baroni 2004 et Baayen 2008) ne procèdent pas à une lemmatisation, contrairement aux méthodes décrites dans (Lu 2014).

Enfin, nous avons essayé d'exploiter les courbes de croissance du vocabulaire. L'objectif est la caractérisation de traits grammaticaux et lexicaux permettant pour chaque enregistrement d'établir une différence de niveau, nous avons donc aussi utilisé les cinq premières valeurs des courbes de croissance du vocabulaire pour chaque transcription, ce qui correspond aux effectifs cumulés des hapax par tranches de 100 occurrences. Nous avons ici restreint la fenêtre de la mesure des hapax, en raison de la petitesse de la taille des donnés. Dans les travaux précédents (Baayen 2001, 2008), le nombre d'hapax est mesuré par incrémentations successives de 500 ou 1000 tokens. Un jeu de données fondé sur l'augmentation des hapax au fil du monologue est donc construit avec cinq effectifs d'hapax pour chaque sujet. On a ensuite procédé à une classification automatique à partir de TiMBL (Daelmans et al. 1994).

2.5 Classifieur retenu

Le classifieur TiMBL, déjà mis à contribution dans plus d'une centaine d'études, permet l'estimation des traits les plus déterminants dans la classification, le calcul du GainRatio (Daelemans et al. 2005) rend possibles une hiérarchisation des traits (*features*) et une optimisation de la classification fondée sur la pondération des traits les plus pertinents dans la classification. C'est ce qui le distingue de la famille des algorithmes k-nn dont il relève.

Pour rappel, le classifieur procède dans un premier temps à un apprentissage en mémorisant les différents vecteurs et leur classe, et en calculant une pondération des variables basée sur la notion d'entropie. Dans un second temps, de nouveaux vecteurs (sans classe) sont présentés au module de classification pour se voir attribuer une classe. Il y a donc un échantillon pour chaque phase. Dans le cas de notre expérience, l'échantillon que nous utilisons ne comprend que 60 observations (converties en 60 vecteurs de variables), ce qui est relativement faible. Afin de pallier ce manque, nous utilisons l'option 'leave-one-out' de TiMBL qui permet d'utiliser le même échantillon pour l'apprentissage et le test. Le programme effectue 60 passages de classification. A chaque passage, 59

vecteurs sont utilisés pour l'apprentissage, le dernier étant utilisé pour le test. Une fois la classification effectuée, des statistiques de classification sont renvoyées, ce qui permet de mesurer la précision de la classification. En outre, la pondération des variables composant les vecteurs est affichée. Cela permet d'avoir une vue précise sur la pondération attribuée à chacune d'entre elles, et ainsi de voir les variables les plus significative du point de vue de leur potentiel de réorganisation des informations. Nous avons donc utilisé l'option *leave-one-out* dans nos classifications.

3. Résultats

A titre de comparaison, rappelons d'abord que la classification obtenue par Anne Tortel sur la base des métriques du rythme était de 69% pour les trois groupes de locuteurs, mais les données portaient sur les productions lues des sujets, non sur leur production spontanée.

3.1 Classification fondée sur la complexité syntaxique

La précision de la classification sur la base des 23 traits retenus pour l'analyse syntaxique est très moyenne (48%) et n'a de sens que pour la classification des locuteurs les moins avancés (FR1)

	FR1	FR2	GB
FR1	**14**	3	3
FR2	6	**8**	6
GB	4	9	**7**

TABLE 2 : Matrice de confusion des trois groupes de locuteurs (toutes métriques confondues)

3.2 Complexité lexicale

La précision est de 48,33% (29/60) et donne la matrice de confusion suivante :

	FR1	FR2	GB
FR1	**12**	5	3
FR2	8	**9**	3
GB	5	7	**8**

TABLE 3 : Matrice de confusion des trois groupes de locuteurs (toutes métriques confondue

TiMBL fait apparaître la pertinence des métriques suivantes : adjv (le rôle des adjectifs) et le type-token-ratio (TTR) brut, ce qui s'explique en partie par la brièveté des monologues. Comme le

rappelle (entre autres) Baayen 2001, la mesure du TTR est problématique en ce qu'elle est sensible à la taille des échantillons. La taille recommandée par (Biber 2012) dans une étude sur le TTR est de 450 tokens. Le rapport quantitatif entre le nombre de types et le nombre de tokens est au cœur des modélisations LNRE (*Large Number of Rare Events*).

3.3 Courbes de l'accroissement du vocabulaire et modèles LNRE

La Figure 1 présente la croissance des types (en ordonnées) en fonction de l'accroissement de l'échantillon (nombre de *tokens* en abscisse). La première série de courbes montre la croissance des hapax (en bas), et la deuxième représente la croissance des types en fonction de l'accroissement de l'échantillon (N). On voit que, sur l'échantillon analysé (tous les locuteurs de chaque groupe confondus), la courbe de croissance ne permet pas de discriminer entre les différents groupes. Les micro-variations (les courbes se croisent) sont également imputables à la variation individuelle des locuteurs analysés, car chaque courbe est constituée à partir des effectifs cumulés des 20 locuteurs.

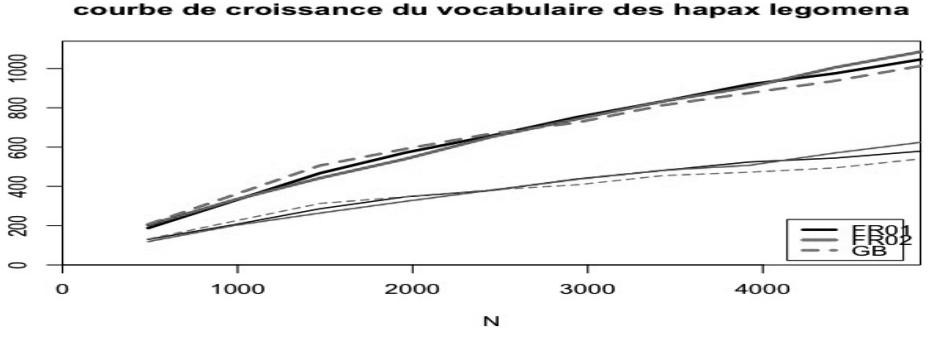


FIGURE 1 : Courbes des *hapax legomena* et du nombre de types en fonction du nombre de tokens

Nous avons également calculé la richesse lexicale, implémentée par Baayen dans le package {languageR} par la fonction compare.richness.fnc(), qui compare les courbes de croissance du vocabulaire (Baayen 2008, Turenne 2016). La comparaison des corpus deux à deux (ramenée, pour les trois sous-corpus, aux 5 027 premiers tokens) ne donne pas de différence significative (p= 0,1969) pour des taux d'accroissement du vocabulaire de 0,11717 pour FR01, de 0,10961 pour GB et de 0.12751 pour FR02. La traduction intuitive que Baayen 2008 donne à cette métrique est que la probabilité que le mot suivant le 5 027^{e} soit une occurrence d'un type inédit jusque là est d'une chance 1 sur 9. Cette fonction est sensible à la taille de l'échantillon, ce qui rend problématique son utilisation avec les corpus d'apprenants, de taille plus restreinte.

4. Discussion et conclusion

Plusieurs ordres de problématiques sont mobilisables, au-delà des problèmes de taille d'échantillon et du traitement statistique choisi. Même avec la méthode du *leave one out*, le nombre de documents pour entraîner le classifieur est très limité. On pourrait affiner les classifications en ne retenant que les traits sélectionnés par TiMBL comme étant davantage pertinents (par exemple le TTR brut), mais on ne saurait rien du caractère *ad hoc* de cette restriction du nombre de métriques pertinentes et on multiplierait les risques d'*overtfitting* sur les données. D'autres méthodes d'analyses

statistiques que la classification étaient possibles pour les données recueillies, en particulier des modèles *stepwise* de régression (cf, entre autres, Johnson 2008).

4.1 Métriques de l'écrit et métriques de l'oral

L'apprenant parle comme il écrit, c'est à ce type de caractéristique qu'on reconnaît qu'il ne parle pas comme un natif. Derrière cette boutade, se dissimulent l'importance de la parataxe en AND chez les natifs et l'importance de l'hypotaxe en SO chez les non-natifs. Les métriques ne rendent pas immédiatement compte des différences, flagrante à l'oral, des constructions prosodiques et de la structuration du discours. Nous y reviendrons et détaillons peu les métriques de complexité syntaxique, le statut de la T-unit pour la caractérisation de l'oral n'est pas idoine, pas plus que notre ponctuation comme pis-aller de la structuration prosodique.

Dans cette approche, les propositions verbales non-finies ne sont pas comptées comme des propositions. Il existe par ailleurs d'autres métriques de complexité, comme le nombre moyen de mots devant le verbe de la principale, le nombre de modificateurs (adjectif ou adverbe) par syntagme nominal, ainsi que des mesures plus techniques fondées sur le corpus arboré (nombre de nœuds dans les phrases). Pour les propositions non-finies, les infinitives et les formes en *-ing* gagneraient à être distinguées, particulièrement pour l'analyse des constructions des francophones.

Sur la base de l'ensemble des métriques mobilisées dans nos analyses (52), la précision obtenue est assez faible (43%). La tâche est plus complexe avec l'ensemble des traits.

	FR1	FR2	GB
FR1	**12**	8	0
FR2	6	**9**	5
GB	5	10	**5**

TABLE 4 : Matrice de confusion des trois groupes de locuteurs (toutes métriques confondues)

4.2 Prolongements LNRE

La distribution de la loi de Zipf vaut également pour les corpus oraux, mais, chez les apprenants, à condition de disposer de suffisamment de corpus, les événements rares (hapax, mots rares) ne seront pas si fréquents que cela et d'autres éléments sont susceptibles d'être répétés.

Nous avons signalé que l'analyse a été conduite par courbe représentant un groupe et n'avons pas randomisé les occurrences attribuables aux différents locuteurs (nous n'avons donc pas neutralisé la variation inter-locuteurs). Une approche des courbes de vocabulaire sujet par sujet, dans une version plus subtile faisant intervenir la randomisation, pourrait donner des résultats différents. Nous n'avons pris en compte la variation inter-locuteur que dans les cinq relevés du nombre d'hapax par tranche de 100 occurrences. Le vocabulaire des apprenants étant plus limité, il nous a paru intéressant d'en mesurer les effets sur la raréfaction des mots nouveaux au fur et à mesure de l'évolution du discours. A titre expérimental, nous avons comptabilisé l'accroissement des hapax

par tranches de 100 tokens. Sur les cent premiers mots, on mesure le nombre d'hapax, sur les tranches suivantes, le différentiel d'hapax entre deux tranches.

La classification sur ces valeurs initiales des courbes d'accroissement du vocabulaire (*option leave one out*) ne donne que 40% de précision (Cf Table 5). La différence de longueur des productions reste l'indice le plus clair de distinction entre groupes de locuteurs, c'est ce qui explique la bonne classification du groupe FR1 : les valeurs cumulées sont plus souvent plus faibles au-delà de 400 mots pour le nombre d'hapax de la dernière tranche d'occurrences, car il y a parfois moins de 400 mots. Contrairement à ce que l'on pourrait penser, les locuteurs natifs n'ont pas une expression plus variée que les non-natifs du groupe FR2, ce qui explique leur très mauvaise classification. La capacité à produire régulièrement des mots nouveaux ne permet pas à elle seule une identification automatique de locuteurs avancé ou natifs. Les locuteurs avancés sont (mal) classés comme des locuteurs moins avancés car ils ont des scores avoisinants. En clair, c'est la différence de *fluency* qui est la plus manifeste entre les groupes de locuteurs, ce que reflète le nombre de mots prononcés dans les deux minutes que dure l'exécution de la consigne.

	FR1	FR2	GB
FR1	**16**	3	1
FR2	17	**2**	1
GB	6	8	**6**

TABLE 5 : Matrice de confusion des trois groupes de locuteurs (méthode des hapax)

Les limites en expression des apprenants pourraient avoir une incidence sur l'occurrence des événements rares au profit de formules récurrentes rassurantes type 'doudous' ('*teddy bear', cf.* Hasselgren 1994 ; Ellis 2012) ou de séquences préfabriquées, « reliability islands » (Dechert 1983). Ce type d'approfondissement gagnerait à envisager l'analyse en n-grams et, surtout, l'analyse de corpus longitudinaux de taille plus conséquente pourrait révéler des courbes de croissance du vocabulaire plus rapidement asymptotiques que pour les natifs. Reste à voir si le discours des apprenants (Gries & Ellis 2015) se laisse simplement régler par les paramètres alpha, A et C (Baayen 1991, Evert & Baroni 2004, Baayen 2008) des modèles LNRE. Le package {zipfR} permettrait d'essayer d'extrapoler les courbes de croissance du vocabulaire pour tester la robustesse des modèles LNRE et sans doute affiner les paramètres alpha, A et C au sein des modèles LNRE pour la modélisation des courbes de croissance selon les groupes de locuteur, voire selon les locuteurs et enfin évaluer la pertinence des trois modélisations possibles à partir de leurs trois implémentations dans le package {zipfr} (lnre.gigp; décrite dans Baayen 2001) Generalized Inverse Gauss Poisson Zipf-Mandelbrot (lnre.zm;Evert, 2004) et finite Zipf-Mandelbrot (lnre.fzm; Evert, 2004). Voici les spectres fréquentiels des trois groupes (toujours en effectifs cumulés, représentés en ordonnée). La première série d'histogrammes à gauche correspond aux hapax, puis aux mots ayant deux occurrences dans le corpus, et ainsi de suite.

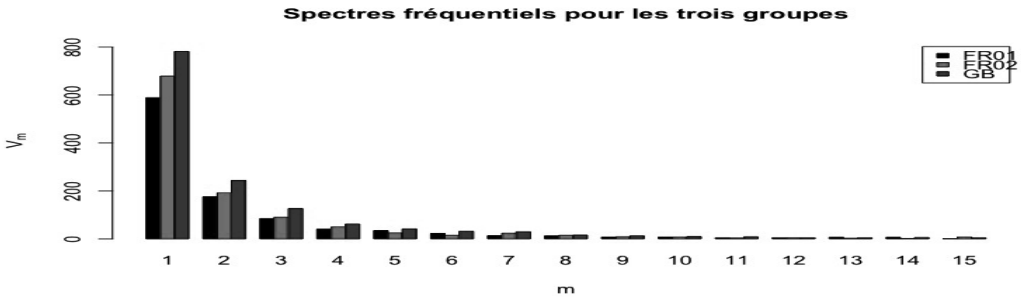


FIGURE 2 : Spectres fréquentiels des trois groupes (effectifs des 20 locuteurs cumulés)

4.3 Approche davantage qualitative : les profils lexicaux (Laufer & Nation 1995)

Il existe une version plus élaborée que les deux mesures ls1 et ls2 pour juger de la rareté des mots employés. Le *Lexical Frequency profile* de lextutor.ca (inspiré de Laufer & Nation 1995) procède par échantillonnage des productions selon quatre catégories de fréquences de référence (inventaire des 1 000 premiers mots les plus fréquents, inventaire des 2 000 mots suivants les plus fréquents, familles lexicales du vocabulaire universitaire et domaine « hors liste »). La composition de chaque production est alors analysée comme l'addition des quatre pourcentages de ces quatre « inventaires » lexicaux. Cette approche davantage qualitative donnerait peut-être de meilleurs résultats, étant donné la nature du corpus. Bien des locuteurs étant des personnels universitaires, le pourcentage de mots « académiques » est conforme au personnel recruté pour les locuteurs du groupe FR1. La fréquence des mots employés semble assez judicieuse en première approximation. En raison de nombreux problèmes de lemmatisation et de mots non-reconnus (*Aix, Aix-en-Provence*), nous n'avons pas pris en compte des résultats obtenus sur l'interface en ligne http://www.lextutor.ca/freq/train/. Les critiques adressées à cette méthodologie dont (Lu 2014) se fait l'écho laissent entendre que la discrimination n'est vraiment opératoire que pour un certain volume de données. On voit la difficulté posée par l'absence de base de données de référence.

4.4 Productions des apprenants à l'oral et fréquences de référence

De manière plus générale, il convient d'établir des inventaires fréquentiels de référence pour les non-natifs. Ce type de recherche devient pressant, dans la mesure où les travaux fondés sur l'usage sont obligés de se rabattre sur les bases de données des fréquences lexicales établies sur des corpus natifs (pour l'anglais, c'est souvent le CELEX, cf. Baayen et al. 1994). Pour l'analyse des productions d'apprenants, le paramètre de la tâche devra être prise en compte : le nombre de locuteurs qui se sont appuyés dans leur monologue sur les éléments du décor est impressionnant, comparés aux natifs, qui ont abordé des sujets plus divers.

Pour les corpus d'apprenants oraux, une difficulté supplémentaire réside dans le statut non-lexical sans doute trop facilement accordé aux pauses pleines. Parmi les points qui nous apparaissent les plus déterminants, les choix de transcrire les pauses par des formes distinctes interrogent à la fois la structuration de l'énoncé, la réalisation des voyelles et le niveau de l'apprenant. Il se trouve que les pauses pleines sont plutôt codées « heu » ou « euh » pour les apprenants FR1, «hum » pour les FR2 et « erm » pour les natifs. Sans même nous livrer à une analyse précise des réalisations phonétiques de leurs traits acoustiques (voir Chlébowski 2015 pour une analyse sur l'anglais de Newcastle), la

qualité de la réalisation de la voyelle diffère. Une observation un peu fine des pauses pleines et de leur transcription textuelle dans le corpus ANGLISH révèle la complexité de ce phénomène. Il y a véritablement toute une réflexion à conduire, tant sur le statut de tokenisation des réalisations phonétiques dans les corpus oraux (Ballier 2016) que sur leur signification (Käger 2016). Ceci dépasse le cadre de cet article, mais on peut songer que les niveaux des apprenants et les strates d'interlangue pourraient également s'analyser en termes de maîtrise de la position des articulateurs, et notamment de la langue (les *articulatory settings* de Catford 1984). On assisterait à une réalisation davantage labialisée par défaut chez les francophones moins avancés (le « *heu* » des FR1), à des réalisations plus proches des natifs en termes de timbre vocalique ou de recours au trait nasal chez les FR2 (le « *hum »),* et à une position de la langue plus en avant et une labialisation moins marquée chez les anglophones (« *erm* »).

L'examen de l'ordre dans la fréquence des occurrences (loi de Zipf) des pauses pleines souligne les problèmes de lemmatisation des pauses pleines dans une perspective interlangue (le statut et la fréquence de *heu, euh, hum*), sans compter les cas de réalisations clitiques (telles *but ahem* réalisé en deux syllabes). *Erm* est au 6e rang des occurrences chez les natifs (après les classiques *the, and, I, to, a*), *hum* est au sixième rang chez les locuteurs avancés FR2 et les transcriptions retenues pour les pauses pleines des FR1 se répartissent entre *hum* (8^{e} rang), *heu* (15^{e} rang) et *euh* (18^{e} rang). Les effectifs cumulés situant les occurrences de pause pleine au troisième rang des tokens des locuteurs FR1. Plutôt que de voir dans cette distribution quasi-complémentaire des pauses (*erm/hum/heu*) des incohérences transcriptionnelles, on peut y voir des variantes sémantiquement significatives des pauses pleines dont le rôle éventuel doit être analysé, des marqueurs du discours dont la contribution à la structuration de la hiérarchie des constituants prosodiques ne sont sans doute pas réductibles à la T-unit. (Paris 2015 :203-5) montre les limites de la notion de T-unit pour l'analyse de l'oral et notamment pour la prise en compte des faux-départs, des incises, du discours rapportés ou de ce qu'elle nomme des « propositions concaténées » dans l'analyse de ses enregistrements. Pour des données orales, le constituant prosodique inter-pauses pleines serait une unité plus pertinente que la T-unit. Manque cruellement une métrique de complexité à l'oral davantage fondée sur la prosodie, quelle que soit son empan dans la hiérarchie prosodique (Nespor & Vogel 2007).

4.5 Effets du genre

Le corpus ANGLISH, étant contrôlé, il y a 10 hommes et 10 femmes dans chaque groupe de locuteurs. Nous avons essayé de voir si les métriques rendaient compte de ces différences entre hommes et femmes. La tâche est plus complexe, mais la matrice de confusion pour les niveaux et les genres donne les résultats suivants, avec l'ensemble des métriques considérées, pour une précision de 23% seulement :

	FFR1	FFR2	FGB	HFR1	HFR2	HGB
FFR1	5	2	0	2	1	0
FFR2	2	4	1	3	0	0
FGB	1	2	0	2	4	1
HFR1	5	4	0	0	1	0

HFR2	0	2	3	1	3	1
HGB	1	2	2	2	1	2

TABLE 6 : Matrice de confusion par niveau et genre des groupes de locuteurs

4.5 Conclusions

Le projet d'ensemble est celui d'une analyse multidimensionnelle des productions orales des apprenants. On cherche, entre autres propriétés, à établir des corrélations ou des distinctions strictes entre les compétences lexicales, syntaxiques et phonologiques des apprenants. La comparaison n'est ici que partielle et a concerné la partie spontanée du corpus, là où la thèse d'Anne Tortel a porté sur les parties lues du corpus ANGLISH. Pour le moment, ce corpus dans sa dimension phonétique n'a été étudié qu'à partir de sa composante lue (Tortel 2008, Ballier et al. 2016), afin de favoriser les comparaisons inter-sujets. À l'inverse, le travail sur la production libre en spontané permet le type de raffinement que nous venons de proposer. Soulignons donc une fois encore l'intérêt de ce corpus ANGLISH et la diversité des réalisations qui s'y donne à voir et à entendre. En particulier, les productions des hommes et des femmes ne sont sans doute pas équivalentes au plan de la prosodie, alors que la classification ne semble pas l'établir aussi nettement.

Le deuxième point à faire valoir porte sur la diversité des technologies maintenant abordables pour l'étude des corpus d'apprenants (Díaz et al. 2013). Comme l'écrit (Lu 2014), il existe maintenant une panoplie d'outils qui constituent un terrain d'entente entre linguistes et programmeurs : "something in the middle ground, something that enables novice language and linguistics researchers to use more sophisticated and powerful corpus annotation and analysis tools than concordancing programs and yet still does not require programming".

En première analyse, sur des échantillons de données assez restreints, les métriques de complexité lexicale et syntaxique ne se substituent pas à des tests initiaux d'évaluations du niveau des apprenants. La taille restreinte du corpus est évidemment un obstacle, ce qui nous a conduits à multiplier les approches quantitatives par métriques et les différentes métriques possibles dans ce qui reste une expérimentation à très petite échelle. La taille restreinte pourrait être un avantage pour conduire également des approches plus qualitatives fondées sur les récurrences des n-grams. Reste que la piste automatique n'est pas complètement satisfaisante, même s'il conviendrait pour la réfuter complètement sur le jeu de données étudié de refaire une classification du niveau relatif des différents locuteurs francophones sur la base d'évaluations d'experts. C'est notre prochaine étape, en plus de la réplication de la méthodologie sur un corpus longitudinal de 135 locuteurs.

Remerciements

Nous remercions chaleureusement Anne Tortel pour la mise à disposition de son corpus, déposé sur le SLDR. Nous remercions Xiaofei Lu pour ses explications et ses démonstrations de ses solutions logicielles et Stefan Evert et Marco Baroni pour la clarté de leur documentation du package {ZipfR} et de son site dédié http://zipfr.r-forge.r-project.org ainsi que, pour leurs commentaires, les participants du workshop du 30 mars (Paris Diderot) et les relecteurs anonymes de TALN.

Références

BALLIER N. MARTIN, PH. (2015). Speech annotation of learner corpora. In: Granger, S., Gilquin, G., Meunier, F., (eds), *The Cambridge Handbook of Learner Corpus Research*, Cambridge: Cambridge University Press.

BALLIER N. (2016). Du dictionnaire lexico-phonétisé aux corpus oraux, quelques problèmes épistémologiques pour l'école de Guierre. *Histoire Epistémologie Langage*, à paraître.

BALLIER N., MARTIN, PH. AMAND, M. (2016). Variabilité des syllabes réalisées par des apprenants de l'anglais, *JEP 2016*, Paris, 8 pages.

BAAYEN R.H. (2001). *Word Frequency Distributions*. Dordrecht, Boston & London: Kluwer.

BAAYEN R.H. (2008). *Analysing Linguistic Data with R*. Cambridge : CUP.

BARONI M. (2009). Distributions in Text. In A. Lüdeling, M. Kytö (eds), *Corpus Linguistics. An International Handbook*, 803–821. Berlin, New York: de Gruyter Mouton.

BARONI M. EVERT, S. (2014). The zipfR package for lexical statistics: A tutorial introduction , zipfR version 0.6-7.

BENTZ C., BUTTERY P. (2014). Towards a Computational Model of Grammaticalization and Lexical Diversity. Proceedings of 5th *Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*, 38-42.

CHITEZ M. (2014). *Learner corpus profiles: The case of Romanian Learner English*. Oxford : Peter Lang.

CHLEBOWSKI A. (2015*). Nasal Grunts » in the NECTE corpus An experimental investigation.* Mémoire de M2, Université Paris Diderot.

DAELEMANS W., ZAVREL, J., VAN DER SLOOT, K., & VAN DEN BOSCH A. (2004). Timbl: Tilburg memory-based learner. *Tilburg University*.

DAELEMANS W., VAN DEN BOSCH A. (2005). *Memory-based Language Processing*. Cambridge: CUP.

DECHERT, H.W., 1983, How a story is done in a second language, in: Farch, C. & Kasper, G. 1983d, Strategies in interlanguage communication, London, Longman., 175-196.

ELLIS, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17-44.

DIAZ-NEGRILLO A., BALLIER N., THOMPSON P. (eds.). (2013). *Automatic treatment and analysis of learner corpus data*. (Studies in Corpus Linguistics 59). Amsterdam: Benjamins.

DRAGER, K. (2016). Constructing style: phonetic variation in quotative and discourse particle *like* In Heike Pichler (ed). *Discourse-Pragmatic Variation and Change in English:New Methods and Insights*, Cambridge : CUP, 232-251.

EVERT, S., & BARONI, M. (2007). zipfR: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, 29-32. Association for Computational Linguistics.

EVERT, S., BARONI, M., (2006). The zipfR package.*http://cran.r-project.org/doc/packages/zipfR.pdf*

GAILLAT, T. (2013). Annotation automatique d'un corpus d'apprenants d'anglais avec un jeu d'étiquettes modifié du Penn Treebank. Actes de *la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, 271–284.

GRIES, S. T., & ELLIS, N. C. (2015). Statistical Measures for Usage‐Based Linguistics. *Language Learning*, *65* (S1), 228-255.

HAWKINS J., BUTTER P. (2010). Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal* 1(01), 1-23.

HAWKINS J. A., FILIPOVIĆ L. (2012). *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*. United Kingdom: Cambridge University Press.

HASSELGREN, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, *4*(2), 237-258.

HUNT K.W. (1965). *Grammatical Structures Written at Three Grade Levels*. Champaign IL: National Council of Teachers of English.

HUNT K.W. (1970). Do Sentences in the Second Language Grow Like Those in the First? *TESOL Quarterly* 4(3), 195–202.

JOHNSON, K. (2008). *Quantitative Methods in Linguistics*. Londres :Blackwell.

KAMEEN P.T. (1979). Syntactic Skill and ESL Writing Quality. In C. Yorio, K. Perkins, J. Schachter (eds), *On TESOL '79: The Learner in Focus*, 343–364. Washington DC: TESOL.

KLEIN, D., & MANNING, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1* (pp. 423-430). Association for Computational Linguistics.

LAUFER, B., et NATION, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16 (3), 307-322

LECLERCQ P., EDMONDS A., HILTON H. (Eds.). (2014). *Measuring L2 Proficiency: Perspectives from SLA*. Bristol: Multilingual Matters.

LU X. (2010). Automatic Analysis of Syntactic Complexity in Second Language Writing. *International Journal of Corpus Linguistics* 15, 474–496.

LU X. (2011). A Corpus-based Evaluation of Syntactic Complexity Measures as Indices of College-level ESL Writers' Language Development. *TESOL Quarterly* 45, 36–62.

LU X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal* 96, 190–208

LU X. (2014). *Computational Methods for Corpus Annotation and Analysis*, Dordrecht: Springer.

LU, X. (2016). L2 Syntactic Complexity Analyzer, http://www.personal.psu.edu/xxl13/download.html, consulté le 20/04/2016

NESPOR, M., & VOGEL, I. (2007). *Prosodic Phonology*, Berlin : Mouton de Gruyter.

OVTCHAROV V., COBB T., HALTER R. (2006). La richesse lexicale des productions orales: mesure fiable du niveau de compétence langagière. *Canadian modern language review* 63(1), 107-125.

PARIS, J. (2015). *Lumière sur le développement de la production de langage non-littéral en L2. Pour une comparaison avec l'acquisition des langues maternelles*, thèse non publiée, Paris 3.

TONO Y. (2013). Automatic Extraction of L2 Criterial Lexicogrammatical Features across Pseudo-longitudinal Learner Corpora: Using Edit Distance and Variability-based Neighbour Clustering. In C. Bardel, C. Lindqvist & B. Laufer (eds.), *L2 Vocabulary Acquisition, Knowledge and Use: New Perspectives on Assessment and Corpus Analysis*, 149–176. (Eurosla Monographs Series 2). The European Second Language Association.

TORTEL A. (2008). ANGLISH: Base de données comparatives de l'anglais lu, répété et parlé en L1 & L2, *Travaux Interdisciplinaires du Laboratoire Parole et Langage (TIPA)* 27, 111-122.

TORTEL, A. (2009). *Evaluation qualitative de la prosodie d'apprenants français: apport de paramétrisations prosodiques*. Thèse de doctorat non publiée. Aix-Marseille University.

TURENNE N. (2006). *Analyse de données textuelles sous R*. Paris : ISTE.

WOLFE-QUINTERO, K., INAGAKI, S. & KIM, H.-Y. 1998. *Second language deveopment in writing: Measures of fluency, accuracy and complexity* [Technical Report 17]. Honolulu: University of Hawaii, Second Language Teaching and Curriculum Center.

ZIPF G.K. (1949) *Human Behavior and the Principle of Least Effort*. Cambridge (Massachusetts): Addison-Wesley.

De l'exemple construit à l'exemple attesté : un système de requêtes syntaxiques pour non-spécialistes

Ilaine Wang[1] Sylvain Kahane[1] Isabelle Tellier[2]
(1) MoDyCo (UMR 7114), CNRS, Université Paris Ouest Nanterre La Défense
(2) LaTTiCe (UMR 8094), CNRS, ENS Paris, Université Sorbonne Nouvelle - Paris 3,
PSL Research University, USPC (Université Sorbonne Paris Cité)
i.wang@u-paris10.fr, sylvain@kahane.fr, isabelle.tellier@univ-paris3.fr

RÉSUMÉ

Notre objectif est de permettre aux apprenants de langue d'accéder aux données présentes dans un corpus en facilitant la formulation de requêtes portant sur des critères syntaxiques. Nous proposons une méthodologie utilisant des mesures de similarité classiques pour comparer des séquences d'étiquettes obtenues par des programmes d'annotation.

ABSTRACT

From built examples to attested examples : a syntax-based query system for non-specialists

Our purpose is to allow non-specialists like language learners to access corpora, making syntactic queries easy to express. We propose a methodology including a syntactic parser and using common similarity measures to compare sequences of (automatically produced) morphosyntactic tags.

MOTS-CLÉS : linguistique de corpus, système de requête, construction syntaxique.

KEYWORDS: corpus linguistics, query system, syntactic construction.

1 Introduction

Depuis l'ère numérique, le développement de la linguistique de corpus s'accompagne du développement de systèmes de requêtes permettant d'exploiter les corpus comme des ressources linguistiques. De la même façon que la recherche d'information s'est dotée de moteurs de recherche, la linguistique de corpus s'est équipée d'outils tels que les concordanciers, qui reposent sur des requêtes à base de mots-clés. L'exploitation d'un corpus n'est pas seulement rendue possible par l'utilisation de concordanciers, elle est également déterminée par ces derniers : ce que l'on tire des corpus dépend fortement des possibilités offertes par les outils qui les exploitent (Anthony, 2013), et reposer sur des mots-clés peut être une contrainte pour qui s'intéresse à certaines constructions complexes et/ou qui n'ont pas de marqueur lexical précis. On considérera par exemple le cas des propositions relatives, marquées non pas par un item lexical mais par la catégorie grammaticale des pronoms relatifs.

Il n'est possible aujourd'hui de rechercher des structures complexes qu'en les décrivant précisément ce qui suppose de connaître le langage de requête et le jeu d'étiquettes du corpus annoté. Ces connaissances, communes en linguistique outillée et en TAL, requièrent des efforts importants de la part de non-spécialistes tels que les apprenants ou enseignants de langue. Dans la suite de l'article, nous commençons par faire le tour des besoins et des outils actuellement disponibles pour l'utilisation

des corpus en didactique des langues. Nous proposons ensuite une chaîne de traitements prenant en considération les difficultés éventuelles de non-spécialistes et comprenant un système de requête fondé sur une notion de similarité syntaxique.

2 Interrogation de corpus

Les apprenants et les enseignants de langues ne sont généralement pas des linguistes. Ils sont rarement initiés aux méthodes de la linguistique outillée ou du TAL alors qu'ils sont de plus en plus nombreux à percevoir l'intérêt d'accéder à des corpus. Après avoir évoqué les tenants et les aboutissants de l'accès à des données attestées en didactique des langues, nous présentons les outils disponibles pour interroger des corpus, en montrant leurs limites, notamment quand la requête porte sur une construction syntaxique.

2.1 Utilisation des corpus en didactique des langues

Les corpus de locuteurs natifs représentent en didactique des langues une ressource intéressante puisqu'ils constituent pour l'enseignant comme pour l'apprenant des ensembles de documents authentiques dans lesquels il est possible d'observer ce qui est considéré comme naturel ou usuel dans la langue cible (voir notamment les travaux de Chambers (2005, 2010) et de Cavalla (2015) pour l'aide à l'écrit en Français Langue Etrangère). Cette exposition à des données authentiques peut être indirecte (distribution en classe de concordances réalisées au préalable par exemple) ou bien être l'aboutissement d'une démarche plus directe. La seconde méthode est particulièrement exploitée dans l'approche que Johns nomme *Data-Driven Learning* (DDL), et qui considère l'apprenant comme un « chercheur dont l'apprentissage devrait être motivé par l'accès à des données linguistiques » [1] (Johns, 1991, p.2). L'apprenant doit être actif dans son apprentissage, être capable de formuler des hypothèses, d'observer et d'analyser des données langagières pour confirmer ou infirmer lui-même ses hypothèses, et enfin d'en formuler de nouvelles au besoin.

Or, dans la pratique, les apprenants peuvent considérer que les bénéfices évidents d'une confrontation directe à des corpus authentiques ne valent pas les efforts fournis ni le temps employé pour apprendre à utiliser de manière appropriée les outils d'exploration des corpus. Boulton (2012) cite en effet parmi les points négatifs soulevés par ses étudiants la complexité de l'interface de requête et donc la nécessité de recevoir une formation spécifique pour pouvoir exploiter au mieux les corpus. C'est à partir du même constat que Falaise *et al.* (2011) proposent un outil d'exploration de corpus arborés avec une interface plus simple, minimaliste (options cachées) et conviviale (interface graphique et non textuelle), qui n'empêche pas des requêtes fines et précises. Si cette simplification de l'interface permet effectivement de réduire de manière significative le temps nécessaire à la maîtrise de l'outil, elle présuppose néanmoins les mêmes connaissances que précédemment de la part de l'utilisateur.

2.2 Méthodes actuelles de requêtage en linguistique de corpus

Une des méthodes les plus courantes en linguistique de corpus est l'utilisation des concordanciers. Ces derniers, de plus de plus utilisés en didactique des langues, incluent tous au moins deux fonctionalités

1. Dans le texte original : *research workers whose learning needs to be driven by access to linguistic data*".

principales qui ont pour unité de base le mot : d'un côté, le calcul de statistiques mettant en évidence les propriétés du texte (ou corpus) étudié (nombre d'occurrences, distribution, collocation etc.), et de l'autre, les concordances KWIC (KeyWord In Context) où l'on retrouve le mot ou la séquence de mots cibles alignés et dans leur contexte originel. On peut remarquer qu'à la différence des requêtes dans les moteurs de recherche, les suites de mots données en entrée à un concordancier sont généralement des n-grams, soit des séquences de mots strictement contiguës, dont l'ordre est préservé. L'implémentation de *skipgrams* (n-grams non contigus) dans des logiciels de concordance est plus rare mais on note qu'il existe des outils de recherche d'unités phraséologiques à visée pédagogique, dont ConcGram et le Lexicoscope pour le français, qui l'autorisent. Ces derniers tiennent compte des variations de position et de dépendance à l'intérieur d'un syntagme grâce à un système prenant en entrée plusieurs mots [2] dits pivots, soit directement donnés par l'utilisateur, soit associés de manière itérative à partir d'un premier pivot (ou deux pour ConcGram) auquel on adjoint jusqu'à quatre mots co-occurrents repérées par l'outil (Cheng *et al.*, 2006; Kraif & Diwersy, 2012).

Il est toutefois possible de se libérer du mot et d'avoir directement recours aux étiquettes morpho-syntaxiques. En effet, l'appariement de deux segments comme "*la personne que je vois*" et "*ce rêve dont tu parles*" qui n'ont aucune unité lexicale commune mais qui, en revanche, partagent la même structure syntaxique, ne peut être réalisé qu'avec le patron "`DET NOM PROREL PRO VERBE`". Pour formuler une telle requête, l'utilisateur doit non seulement connaître ce jeu d'étiquettes mais également avoir suffisamment de connaissances linguistiques, notamment pour pouvoir associer à un mot la bonne partie du discours. Les expressions régulières permettent une expressivité encore plus grande, mais au prix d'une initiation encore plus poussée. L'outil GrETEL (Augustinus *et al.*, 2012) résoud en partie le problème puisqu'il offre la possibilité d'interroger un treebank en transformant automatiquement un exemple de structure syntaxique en requête, à l'instar de ce que nous proposons. Il permet ainsi à ses utilisateurs de s'affranchir de l'apprentissage d'un langage de requête complexe, mais s'adresse toutefois bien à des linguistes conscients de ce qu'ils recherchent et donc capables de paramétrer la requête en ce sens.

Nous nous attachons à aller plus loin dans l'ouverture des outils d'exploration de corpus en proposant une chaîne de traitements qui comprenne à la fois (1) la réduction de la complexité de l'interface du système de requêtes et (2) la réduction de la profondeur et de la variété des connaissances sollicitées de la part de l'utilisateur. Pour notre problématique, nous avons choisi pour le moment de n'exploiter que les parties du discours, sans prendre en compte la structure arborescente des *treebanks*.

3 Méthodologie

3.1 Chaîne de traitements

Notre objectif étant de simplifier au maximum la tâche de la formulation de la requête pour le non-spécialiste, nous proposons une méthodologie qui permettrait de passer "directement" d'un exemple donné en langage naturel à d'autres exemples illustrant la même construction syntaxique. Toutes les étapes de transformation et de comparaison des données seraient assumées par des traitements automatisés et ne solliciteraient donc pas plus de connaissances que celles nécessaires à la validation (ou l'invalidation) des résultats donnés en sortie. La chaîne de traitements complète détaillée en

2. Par mot, on entend ici le mot tel quel (forme fléchie) ou bien le lemme qui lui correspond, permettant aux utilisateurs de considérer ou non les variations morphologiques.

Figure 1 avec un exemple de proposition relative s'articule autour des étapes suivantes :

1. l'analyse (morpho)syntaxique automatique du (ou des) segment(s) donné(s) en entrée par l'utilisateur [3] ;
2. la transformation de l'input en langage naturel en une requête interprétable par la machine ;
3. la mesure de similarité syntaxique entre la requête et les phrases du corpus ;
4. la proposition à l'utilisateur des segments similaires regroupés en clusters ;
5. la sélection par l'utilisateur de l'exemple qui lui paraît le plus proche de sa requête, permettant ainsi d'affiner la requête initiale ;
6. la proposition en sortie des segments qui appartiennent au cluster choisi.

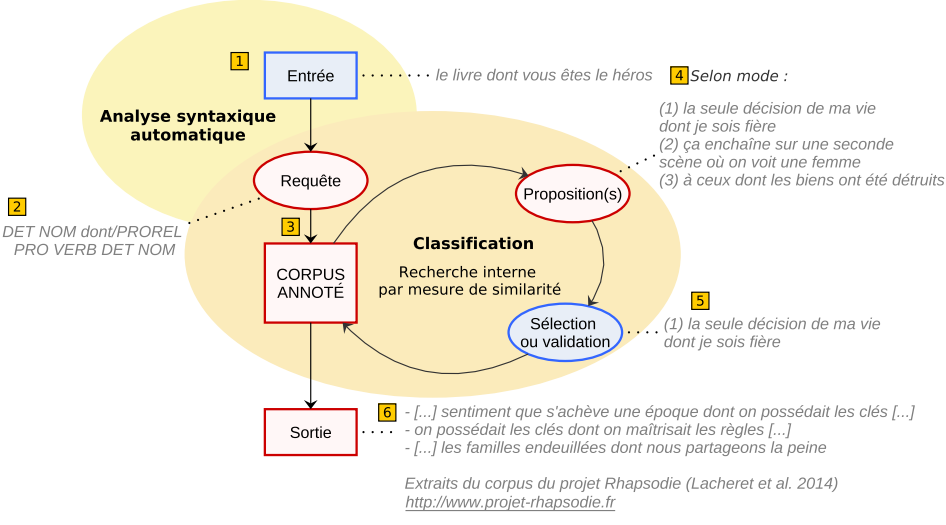


FIGURE 1 – Schéma du système de requêtes syntaxiques envisagé

S'agissant d'un projet en cours de développement, nous nous concentrons dans cette communication sur les trois premières étapes de notre chaîne.

3.2 La similarité comme méthode de recherche souple

Nous avons vu avec l'exemple des propositions relatives que la similarité syntaxique ne pouvait pas reposer uniquement sur une (suite d') unité(s) lexicale(s) mais devrait plutôt être décrite à l'aide de *patrons syntaxiques* sous la forme de séquences d'étiquettes, éventuellement mêlées à des mots. L'idée est bien entendu de pouvoir "matcher" des instances d'une même construction syntaxique en tolérant une certaine variation dans le lexique mais aussi dans la structure elle-même. En effet, si on regarde les propositions types de la Figure 1, on remarque que le premier segment donné par l'outil

3. Les étapes où l'utilisateur doit intervenir sont représentées par des items contournés de bleu et ont été réduites au strict nécessaire afin de rester en accord avec notre objectif de simplification.

ne correspond pas strictement à la requête. Pourtant, alors que leurs étiquettes diffèrent légèrement (on a respectivement "`[...] DET NOM PROREL PRO VERB ADJ`" pour la proposition et "`DET NOM PROREL PRO VERB DET NOM`" pour la requête), la proposition reste pertinente.

Puisqu'il n'est pas évident pour un utilisateur non-spécialiste de définir un patron efficace, c'est-à-dire avec un seuil de tolérance suffisamment élevé pour accepter les variations mais suffisamment bas pour ne pas manquer en précision, nous proposons une méthode basée sur la mesure d'une similarité entre le patron et les segments du corpus. Cette méthodologie a l'avantage d'être plus souple qu'une requête avec des expressions régulières et de permettre de rester plus proche des données en respectant l'approche bottom-up suggérée par le *Data-Driven Learning* [4]. Cette souplesse permet également à l'utilisateur d'avoir le choix entre plusieurs options :

1. faire une recherche en gardant le même mot ou certains mots de l'entrée (plutôt les mots grammaticaux : cas de la première proposition de l'étape 4 dans le schéma) ;
2. faire une recherche et récupérer des structures proches mais qui ne comportent pas forcément les mêmes mots (proposition 2, avec "où" au lieu de "dont") ;
3. si on dispose de ressources lexicales, faire une recherche avec un mot (cette fois-ci plutôt lexical) sémantiquement proche est également une possibilité.

La deuxième option correspond typiquement à la recherche de structures telles que les propositions relatives, puisqu'elles contiennent en français nécessairement une catégorie essentielle, celle du pronom relatif qui peut avoir différentes formes en surface parmi une liste finie. L'outil doit donc être capable d'identifier la catégorie des pronoms relatifs mais ne pas chercher forcément le même pronom que dans l'entrée, et surtout autoriser des variations dans les étiquettes périphériques étant donné que le contexte syntaxique peut être très différent selon la fonction du pronom dans la principale et la subordonnée. La première, quant à elle, se rapproche de ce que propose un concordancier, à la différence que le contexte doit être proche de celui de l'entrée, tandis que la troisième option intègrerait la possibilité d'étendre la requête en se servant de la similarité sémantique, comme c'est déjà le cas pour certaines applications en recherche d'information (moteurs de recherche, systèmes question-réponse) où le(s) mot(s)-clé(s) peuvent être remplacés par des synonymes ou hyperonymes.

Le choix entre ces différentes options pourrait être déterminé par l'utilisateur dès le départ s'il est suffisamment conscient de ce qu'il cherche et suffisamment compétent pour l'identifier. Dans le cas contraire, l'utilisateur pourra déterminer l'option qui lui convient le mieux grâce à la présentation d'un exemple similaire concret issu de chacune des options (voir étape 4 de la chaîne de traitements).

3.3 Mesures de similarité

Nous avons choisi d'utiliser les coefficients de Jaccard et de Dice, largement employées en TAL pour mesurer la similarité, en particulier entre deux mots ou deux chaînes de caractères. Dans notre contexte, il s'agit de mesurer la similarité entre des unités plus larges, des séquences d'étiquettes (`DET NOM PROREL...`) et/ou d'étiquettes couplées avec leur unité lexicale (dont/`PROREL`). Nous explorons également la piste de la distance d'édition (ou distance de Levenshtein), permettant d'évaluer indirectement une similarité. Si la similarité est maximale, la distance est nulle, et vice-versa. Cette alternative est particulièrement intéressante puisque la distance d'édition entre deux

4. Par opposition au *data-based* (littéralement "*basées sur les données*"), les méthodes dites *data-driven* (littéralement "*conduites par les données*") suivent un raisonnement inductif et partent de l'observation des régularités dans les données pour formuler des hypothèses ou les modifier.

"mots" (ou, similairement, entre deux séquences d'étiquettes) M et N se définit par le coût minimal nécessaire pour passer de M à N en effectuant des insertions, ajouts ou substitutions d'unités. Il est par ailleurs possible de pondérer le coût de chaque opération afin d'adapter le calcul de la distance d'édition à nos données. On pourrait en effet tout à fait considérer la suppresion d'un adjectif comme moins coûteux que la suppression d'un verbe ou d'une conjonction.

4 Expériences préliminaires

Des expérimentations sur le coréen langue étrangère sont actuellement réalisées, simulant notamment les recherches d'un apprenant qui éprouverait des difficultés à comprendre les contextes d'usage d'une structure grammaticale donnée et qui aurait besoin de davantage d'inputs authentiques en contexte (Wang, 2016). Nous donnons alors en entrée au programme des phrases qui sont typiquement à disposition des apprenants, celles qui servent à illustrer les explications grammaticales tirées de manuels de langue universitaires [5] et cherchons à comparer la structure de ces phrases à celles du Corpus Sejong (Kim *et al.*, 2007), le corpus de référence pour la langue coréenne. Les tests sont effectués sur le corpus annoté en partie du discours uniquement (environ 13,5 millions de tokens) et constitué d'échantillons de langue variés, écrits comme oraux.

5 Conclusion et perspectives

Nous avons vu qu'au coeur de notre étude se trouvait la simplification de l'accès aux corpus annotés pour un public non spécialiste, et bien que certaines études défendent que la confrontation à des données authentiques est bénéfique à un stade précoce de l'apprentissage (Holec, 1990; Boulton, 2009) la question de l'autonomie de l'apprenant face à la complexité des données authentiques se pose assurément, d'autant plus que nous avons fait le choix de travailler uniquement avec des corpus monolingues. Plusieurs options seront étudiées pour appréhender cette difficulté, à la fois au niveau de la présélection des données dès l'entrée (genre des textes et degré de lisibilité notamment) comme il est possible de le faire pour un grand nombre d'outils à des fins didactiques ou non, mais aussi au niveau de la visualisation des données en sortie (coloration syntaxique similaire à ce qui est proposé pour FipsColor (Nebhi *et al.*, 2010) ou encore dictionnaire intégré pour éviter que le vocabulaire n'ajoute une difficulté cognitive supplémentaire à l'analyse des résultats).

Un certain nombre de traitements sont envisagés sur le corpus, dont le regroupement préalable des phrases du corpus en clusters syntaxiques, améliorant ainsi la rapidité de l'outil puisqu'un seul membre représentatif de chaque cluster pourrait être comparé à la requête puis présenté à l'utilisateur. Cette étape supplémentaire évite à ce dernier d'être submergé de données non classées et devoir faire le tri parmi des dizaines voire centaines de résultats comme c'est souvent le cas avec un concordancier.

Notre outil n'est pas à proprement parler pédagogique en lui-même mais nous pensons que ce programme pourrait à terme compléter les ressources didactiques existantes en permettant une focalisation originale sur les structures grammaticales de la langue cible.

5. En l'occurrence, il s'agit des manuels des niveaux 1, 2 et 3 (équivalent à une à trois années d'études en coréen) de l'université de Yonsei et de ceux du centre linguistique de l'Université de Ewha.

Références

ANTHONY L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, **30**(2), 141–161.

AUGUSTINUS L., VANDEGHINSTE V. & VAN EYNDE F. (2012). Example-based treebank querying. In *Proceedings of eighth international conference on Language Resources and Evaluation (LREC'2012)*, p. 3161–3167.

BOULTON A. (2009). Testing the limits of data-driven learning : language proficiency and training. *ReCALL*, **21**(1), 37–54.

BOULTON A. (2012). Beyond concordancing : Multiple affordances of corpora in university language degrees. *Procedia-Social and Behavioral Sciences*, **34**, 33–38.

CAVALLA C. (2015). Collocations transdisciplinaire : réflexion pour l'enseignement. In *Le problème de l'emploi actif et / ou de connaissances passives des phrasèmes chez les apprenants de langues étrangères*. E.M.E & Intercommunication.

CHAMBERS A. (2005). Integrating corpus consultation in language studies. *Language learning & technology*, **9**(2), 111–125.

CHAMBERS A. (2010). L'apprentissage de l'écriture en langue seconde à l'aide d'un corpus spécialisé. *Revue française de linguistique appliquée*, **XV**, 9–20.

CHENG W., GREAVES C. & WARREN M. (2006). From n-gram to skipgram to concgram. *International journal of corpus linguistics*, **11**(4), 411–433.

FALAISE A., TUTIN A. & KRAIF O. (2011). Exploitation d'un corpus arboré pour non spécialistes par des requêtes guidées et des requêtes sémantiques. In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2011)*, Montpellier, France.

HOLEC H. (1990). Des documents authentiques, pour quoi faire. *Mélanges Crapel*, **20**, 65–74.

JOHNS T. (1991). Should you be persuaded : Two samples of data-driven learning materials. *Classroom Concordancing : English Language Research Journal*, **4**, 1–16.

KIM H.-G., KANG B.-M. & HONG J. (2007). 21st Century Sejong Corpora (to be) Completed. *The Korean Language in America*, **12**, 31–42.

KRAIF O. & DIWERSY S. (2012). Le Lexicoscope : un outil pour l'étude de profls combinatoires et l'extraction de constructions lexico-syntaxiques. In *Actes de la 19e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2012)*, p. 399–406.

NEBHI K., GOLDMAN J.-P. & LAENZLINGER C. (2010). FipsColor : grammaire en couleur interactive pour l'apprentissage du français. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2010)*, Montréal, Canada.

WANG I. (2016). A syntax-based query system adapted to language learning and teaching. In *American Association for Corpus Linguistics (AACL) and Technology for Second Language Learning (TSLL) Conference*, Ames, USA : Iowa State University. Poster presentation.

D'un corpus à l'identification automatique d'erreurs d'apprenants

Marie-Paule Jacques
Lidilem, Université Grenoble-Alpes, Bâtiment Stendhal, 38058 Grenoble, France
marie-paule.jacques@univ-grenoble-alpes.fr

RÉSUMÉ

Nous présentons ici une étude préliminaire (work in progress) à l'élaboration d'un système dédié au repérage de zones potentielles d'erreurs dans des textes d'apprenants. Ce repérage permettra d'enrichir un corpus déjà constitué (que nous nommons Corpus de Littéracie Avancé) par un balisage des erreurs. Nous exposons ici la démarche adoptée pour mettre en place ce balisage : un appui sur certains textes du corpus qui sont commentés par les enseignants-correcteurs pour accéder directement aux passages problématiques, puis l'élaboration de requêtes formalisant les écarts à la norme repérés manuellement. Un exemple-jouet illustre la démarche.

ABSTRACT

From Learner corpus to Automatic Error Annotation
I present here a work in progress aiming at designing a tool to automatically retrieve errors in a learner corpus. The corpus is already constituted and available as "Corpus de Littéracie Avancée" in different formats, including an xml TEI-compliant one. My goal is to enrich the xml files with an error tagging. The paper illustrates the way I proceeded to carry out the tagging: I relied on annotations written in the original files by the readers of the texts (the teachers who ordered the work) to directly access to errors. From them, I built relevant queries to search the corpus for similar patterns. An example shows my approach.

MOTS-CLÉS : corpus de textes d'apprenants, repérage des erreurs

KEYWORDS: learner corpus, automatic error identification

1 Introduction

Le travail que nous présentons ici constitue une étude préliminaire à l'élaboration d'un système dédié au repérage automatique de zones potentielles d'erreurs dans les textes d'apprenants que nous avons rassemblés en corpus. Notre objectif à long terme est d'enrichir ce corpus, déjà disponible librement au format xml, d'un balisage des erreurs afin de permettre, d'une part, des recherches sur les particularités de ces écrits, d'autre part, l'élaboration de cours et d'exercices ciblés pour répondre aux besoins manifestés par ces erreurs.

La particularité de notre corpus et de notre étude réside dans le fait qu'il ne s'agit pas ici d'enseignement d'une langue étrangère : les apprenants sont des étudiants massivement français natifs, qui se destinent même pour certains aux métiers de l'enseignement. Au niveau universitaire,

ils ne reçoivent donc plus d'enseignements de langue similaires à ceux que suivent des apprenants de langue étrangère. Cependant, leurs productions manifestent de fréquents écarts à la norme, sur le plan orthographique en premier lieu, mais aussi sur le plan syntaxique et sur le plan textuel. Or, à notre connaissance, il existe peu d'études systématiques à ce palier de l'enseignement.

Le corpus que nous avons bâti se veut une ressource pour de telles études. Nous renvoyons à (Jacques et Rinck, à paraitre) pour sa description complète, nous en rappellerons ici les principaux objectifs pour nous focaliser surtout sur l'enrichissement que nous visons. Nous commençons par une brève revue de travaux antérieurs en lien avec notre étude avant d'expliciter notre démarche et d'indiquer quelques résultats.

2 Corpus d'apprenants pour l'enseignement

2.1 De langue étrangère

La constitution de corpus d'apprenants n'est pas une idée novatrice dans le domaine de l'enseignement des langues étrangères[1]. Ces corpus sont en effet une précieuse source d'informations sur les difficultés et écueils rencontrés lors de l'acquisition d'une langue étrangère : les erreurs qui s'y manifestent sont autant de pistes pour les remédiations ultérieures (Granger *et al.*, 2015). La production des apprenants témoigne de leur degré d'appropriation de la langue cible et de ses particularités linguistiques, aussi bien en terme de maitrise du lexique et de son usage, qu'en terme de syntaxe. La nature de certaines erreurs rend envisageable un repérage automatique : pour nombre d'apprenants d'une langue étrangère, ce sont le choix et l'usage du « bon » item lexical ainsi que l'association des verbes avec les préposition ou types d'arguments corrects qui cristallisent les erreurs de langue. Gaillat (2013) envisage par exemple d'utiliser une annotation en POS comme révélateur de mésusages des formes *this* et *that,* en procédant par comparaison avec un corpus arboré en anglais.

2.2 De langue maternelle

En ce qui concerne les locuteurs natifs et l'apprentissage de la langue maternelle, la situation est autre, du moins en France. La langue ne fait plus l'objet d'un enseignement formel et explicite au-delà du lycée, les quelques corpus d'apprenants de langue maternelle concernent donc surtout les élèves de primaire et de collège (Elalouf *et al.*, 2007). Toutefois, l'idée de s'appuyer sur les erreurs ou fragilités repérées dans ces corpus pour construire l'enseignement est aussi défendue (Cappeau & Roubaud, 2005).

La réalité des productions d'étudiants dans l'enseignement supérieur montre la nécessité de ne pas considérer que tout serait acquis à ce niveau d'études et qu'il ne serait pas nécessaire de continuer l'outillage des apprenants, alors même qu'ils sont confrontés à de nouvelles exigences et à des genres discursifs encore peu familiers, voire inconnus (on n'écrit pas de mémoire de recherche au

[1] Le site de l'Université Catholique de Louvain en recense un certain nombre : https://www.uclouvain.be/473700.html

collège ni même au lycée). C'est pourquoi nous avons depuis plusieurs années entrepris la constitution d'un corpus de textes produits par des étudiants en réponse à une « commande » universitaire, de la licence au master. Ce corpus de « Littéracie avancée », librement accessible, est disponible notamment au format xml proposé par la TEI. L'intérêt de ce format réside dans le fait qu'il comporte diverses balises utiles pour représenter les propriétés que nous jugeons pertinentes, telles que la structure textuelle (découpage en paragraphes, en sections...), et pour indiquer des métadonnées telles que la consigne d'écriture du texte ou l'année d'étude de l'étudiant (pour plus de détails, cf. Jacques & Rinck, à paraître). Le système de balises est en outre particulièrement approprié au repérage de zones remarquables, c'est-à-dire, pour notre propos, de zones dans lesquelles se donnent à voir les besoins de progression des étudiants. Nous voulons donc à terme insérer dans les textes des balises qui repèrent les erreurs, à peu près sur le modèle suivant, où la balise « lexError » signale une erreur de choix lexical (*vêtir* au lieu de *revêtir*) :

<sent> Le langage offre un panel d'inventions infinies, capable de <lexError>vêtir</lexError> différents visages en fonction de ces associations.</sent>

Ce corpus a déjà fait l'objet d'une exploitation pour la construction d'exercices ciblés (Jacques & Rinck, 2015), que nous décrivons succinctement afin de donner corps aux enrichissements visés.

2.3 Du corpus à l'enseignement

Nous avons utilisé une partie seulement du corpus (on comprendra plus loin pourquoi) pour construire des exercices destinés à aider nos étudiants (natifs, rappelons-le) sur deux plans : une amélioration de la maitrise des normes de l'écrit (tournures de phrases, enchainements discursifs, lexique…) et l'adoption d'une posture réflexive sur la langue et sur l'écriture (certains sont de futurs enseignants, ils se doivent d'entrer dans une réflexion métalinguistique). Divers exercices sont ainsi constitués d'extraits – fautifs – du corpus et de réécritures visant un rapprochement avec la norme, à partir desquels il est demandé de sélectionner soit les énoncés incorrects, soit les énoncés corrects. La Figure 1 donne un exemple d'exercice sur l'utilisation de « comment ».

La collecte des extraits susceptibles de servir de supports pour un tel entraînement a été effectuée totalement manuellement, ce qui est une tâche particulièrement gourmande en temps, c'est pourquoi elle n'a pas porté sur la totalité des 338 textes du corpus – dont certains sont des mémoires de 30 à 40 pages – mais sur un sous-ensemble d'un même genre de textes de 2 à 3 pages chacun.

Pour une réelle exploitation du corpus, l'idéal serait un inventaire exhaustif des erreurs, associé à un balisage. Pour cette tâche, le TAL est un moyen d'éviter la lecture extensive des 338 textes et 1,2 millions de mots du corpus : nous le pensons comme une assistance pour accéder aux passages des textes potentiellement erronés. Cet objectif pose de façon assez évidente la question d'une caractérisation à priori de ce qui, du point de vue d'un système de traitement automatique, peut alors « signaler » une erreur.

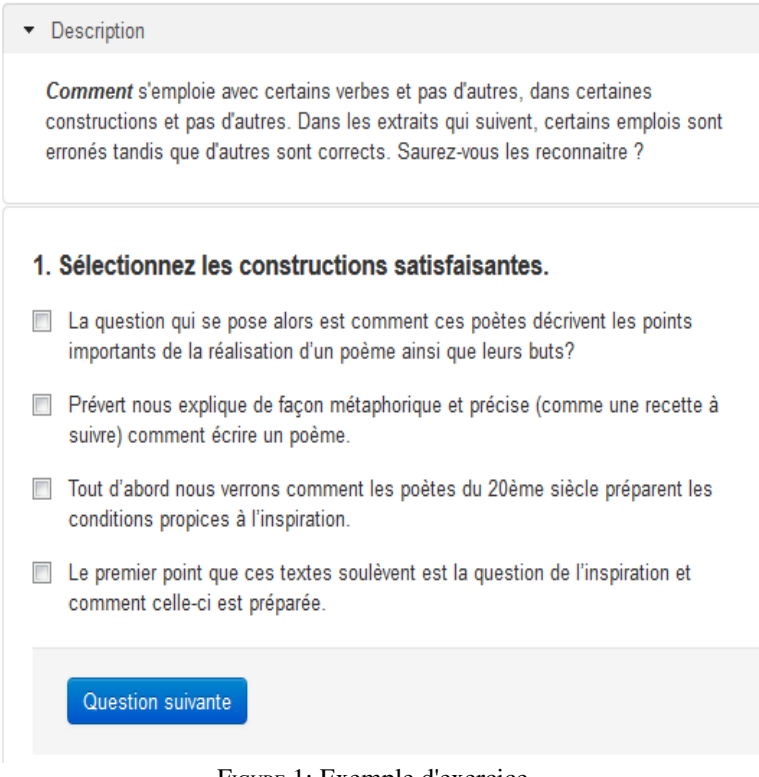


FIGURE 1: Exemple d'exercice

3 Repérer et baliser l'erreur

Reprécisons pour éviter toute ambiguïté que l'objectif ultime est d'insérer dans les textes au format xml des balises délimitant des zones d'erreur et non de proposer une correction automatique des textes. Et ajoutons que, dans la mesure où ce corpus est voué à être ensuite utilisé par des humains et non des machines, nous ne poserons pas d'exigence drastique en termes de délimitation de la zone à baliser : il suffit qu'un repère soit posé de façon à permettre un accès ciblé aux passages des textes potentiellement problématiques. Par exemple, dans l'extrait qui suit, toute la première partie de la phrase est correcte syntaxiquement, on pourrait considérer soit que l'erreur se cristallise sur le point d'interrogation qui n'a pas lieu d'être dans une interrogative indirecte, soit qu'elle tient à l'absence de ponctuation introduisant une interrogative directe, alternative qui est mise en évidence par les réécritures possibles :

Phrase incorrecte : *A partir de cette définition, se pose la question de savoir comment mettre en place cette acculturation à l'écrit en maternelle ?*

Réécriture 1 : *A partir de cette définition, se pose la question de savoir comment mettre en place cette acculturation à l'écrit en maternelle.*

Réécriture 2 : *A partir de cette définition, se pose la question : comment mettre en place cette acculturation à l'écrit en maternelle ?*

Contrairement à Lüdeling *et al* (2005), nous n'avons pas pour objectif de fournir les différentes hypothèses de réécriture – ce seront les recherches menées sur le corpus qui éventuellement inclineront à une hypothèse plutôt qu'à une autre – nous voulons juste signaler la phrase comme comportant un écart à la norme.

Pour construire notre système de traitement, nous procédons en trois temps :

- recueil manuel de passages d'erreurs ;
- classement en types d'erreurs ;
- modélisation en termes de « marqueurs » afin de construire le repérage automatique.

3.1 Des indices : les remarques des correcteurs

Dans le processus de découverte de ces marqueurs, les textes eux-mêmes recèlent des indices : 80 fichiers – fournis originellement dans un format de traitement de texte – comportent des annotations de correcteurs, insérées grâce à la fonction « commentaire » du traitement de texte. Ces annotations sont autant de pointeurs vers les passages problématiques. Elles nous facilitent donc une collecte de ces passages en vue de la typologie et de la modélisation ultérieure.

Une macro dans le traitement de textes nous permet d'extraire chaque passage avec le numéro de la page à laquelle il se trouve dans le document original et le commentaire du correcteur associé. Pour l'heure, nous avons traité 30 fichiers, ce qui nous fournit plus de 750 commentaires différents. Afin d'en extrapoler des données exploitables, il convient de les organiser, au moins grossièrement, en une typologie qui permettra ensuite d'atteindre les régularités.

3.2 Une ébauche de typologie

L'objectif ici est de ranger les passages extraits dans des classes pour spécifier des traitements automatiques différents selon les classes. Nous faisons l'hypothèse en effet que la nature des erreurs induit des procédures de repérage différentes : on ne mettra pas en œuvre la même stratégie pour identifier un problème d'orthographe et un problème de combinatoire verbe / préposition.

En l'état actuel, nos classes sont grossières mais répondent à ce besoin de sérier les extraits. Nous distinguons :

- l'orthographe : les écrits manifestent essentiellement des problèmes d'accords ;
- la formulation : sont ici concernés les passages dans lesquels le correcteur attire l'attention sur une défaillance de la formulation qui peut être liée à un choix de mot(s) inapproprié(s), une construction syntaxique contradictoire avec la combinatoire du lexique choisi, une construction syntaxique bancale qui n'assure pas la complétude de l'énoncé...
- l'agencement textuel (en lien avec la rhétorique du texte) : est ici en jeu le plan du texte et notamment la gestion des enchainements, la structuration, l'ordonnancement des idées...

Dans la mesure où le travail actuel est encore « en chantier », nous n'entrons pas davantage dans cette typologie qui doit encore s'affiner, sinon pour préciser que nous faisons le choix fort de ne pas intégrer cette typologie telle quelle dans le corpus : elle est à l'heure actuelle une heuristique pour les traitements, sa forme définitive sera étroitement liée à notre cadre théorique et aux applications envisagées, elle sera donc en ce sens éminemment contextualisée et probablement peu opératoire pour des chercheurs qui auraient d'autres visées. Elle est un temps de la démarche qui doit permettre de passer de cet ensemble d'occurrences d'erreurs déjà repérées à un système qui permette de délimiter des erreurs sur la majorité des textes du corpus qui n'est pas annotée par les enseignants-correcteurs.

Pour illustrer la démarche, nous allons montrer comment réfléchir le traitement automatique d'un problème du type « formulation », qui met plus particulièrement en jeu la gestion de contraintes syntaxiques au niveau de l'ordre des mots et de la ponctuation.

3.3 Un exemple jouet : le cas des interrogatives indirectes

3.3.1 Exposé du problème

La forme dite « interrogative indirecte » est volontiers présente dans les textes d'étudiants tels que dossiers, synthèses d'articles, mémoires, travaux d'étude et de recherche (TER), car elle y remplit la fonction souvent cruciale de l'expression de la problématique ou de la question traitée. Or elle se montre pour le scripteur d'un maniement peu aisée en raison de ses caractéristiques linguistiques : emploi d'un subordonnant exprimant l'interrogation SANS la forme syntaxique typique de l'interrogation qui elle-même repose sur l'inversion ou le redoublement du sujet et sur l'emploi d'un point d'interrogation. Une formulation normée est : « Nous verrons comment les poètes du 20ème siècle préparent les conditions propices à l'inspiration. » Mais la présence du subordonnant à valeur interrogative entraine régulièrement les étudiants dans une construction « entre-deux », c'est-à-dire qui mélange de façon erronée interrogations directe et indirecte : « On peut se demander comment les poètes de ces deux styles artistiques ont-ils défini leur travail et leur art ? ». On voit que coexistent ici des éléments linguistiques contradictoires : à la fois un subordonnant et une ponctuation interrogative. C'est cette coexistence qui « marque » potentiellement l'erreur de construction.

3.3.2 Traitement et premiers résultats

Notre objectif pour cet exemple jouet est d'identifier automatiquement les erreurs de formulation des interrogatives indirectes. Le corpus a été étiqueté morphosyntaxiquement avec Melt (Denis & Sagot, 2009), nous utilisons TXM (Heiden, 2010) pour élaborer et tester les requêtes qui doivent nous permettre d'atteindre les zones d'erreurs. Une requête représente la traduction en CQL (Corpus Query Language) des caractéristiques linguistiques identifiées manuellement, ici la discordance entre un verbe ou une tournure verbale dédiés à l'interrogation indirecte, par ex. « je me demande, on peut se demander, la question qui se pose... », et des éléments typiques de l'interrogation directe tels qu'une ponctuation finale sous forme d'un point d'interrogation ou une inversion du sujet. Voici un exemple de requête élaborée en CQL, qui correspond aux contraintes suivantes : rechercher un

pronom clitique, puis un mot dont le lemme commence par *demand* ou *question* ou *interr*, ou est *savoir*, puis une conjonction de subordination ou un mot interrogatif qui ne soit pas précédé d'un guillemet ou des deux-points, tout cela dans la limite de la phrase :

[lemma="cl."][]{0,10}[lemma="demand." | lemma="question.*" | lemma="savoir" | lemma="interr.*" | lemma="montr.*"][]{0,15}[lemma!=":" & lemma!="«"][pos="CS" | pos=".*WH"][]{0,15}[lemma="\?"] within sent*

Nous avons ainsi testé trois requêtes appuyées sur des propriétés linguistiques différentes pour le repérage des erreurs de formulation de l'interrogation indirecte. Elles donnent un score de 67 % de rappel et 40 % de précision pour une cinquantaine de contextes renvoyés. Dans la mesure où nous envisageons une validation manuelle des résultats des traitements automatiques, une précision de l'ordre de 40 % nous semble acceptable mais il sera nécessaire d'améliorer le rappel, en veillant toutefois à ne pas dégrader la précision.

4 Conclusion

Nous avons présenté la démarche que nous élaborons pour l'enrichissement d'un corpus existant, consistant en un balisage d'erreurs. Notre approche sera de fournir une ressource neutre quant à l'analyse précise des erreurs, qui est du ressort des études que ce corpus veut servir. L'étape actuelle consiste à bâtir un système à base de TAL qui repère les zones d'erreurs dans les textes et qui est vu comme un auxiliaire qui nous permet de réduire la tâche. Nous posons comme préalable une analyse des caractéristiques formelles de ces erreurs qui constitue alors le point de départ de requêtes en CQL. Même si les résultats ne montrent pas une précision élevée, notre approche permettra de cibler certains passages des textes et donc d'éviter une lecture extensive.

Remerciements

Nous remercions les relecteurs de l'atelier pour leurs remarques sur notre texte.

Références

Cappeau P., Roubaud M.-N. (2005). *Enseigner les outils de la langue avec les productions d'élèves*. Paris : Bordas.

Denis P., Sagot B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. Actes de *PACLIC 2009*.

Elalouf M.-L., Boré C. (2007). Construction et exploitation de corpus d'écrits scolaires. *Revue française de linguistique appliquée* 1/2007 (Vol. XII), 53-70.

Gaillat T. (2013). Annotation automatique d'un corpus d'apprenants d'anglais avec un jeu d'étiquettes modifié du Penn Treebank. Actes de *20e conférence sur le Traitement Automatique des Langues Naturelles*, 271-284.

GRANGER S., GILQUIN G., MEUNIER F. (2015). *The Cambridge Handbook of Learner Corpus Research.* Cambridge: Cambridge University Press.

HEIDEN S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. Actes de *24th Pacific Asia Conference on Language, Information and Computation*.

JACQUES M.-P., RINCK F. (2015). Une linguistique fondamentale et appliquée à base de corpus. *Colloque TRELA, Terrains de Recherche en Linguistique Appliquée.*

JACQUES M.-P., RINCK F. (à paraître). Un corpus de "littéracie avancée" : résultat et point de départ, *Corpus*, numéro spécial sur les corpus d'écrits scolaires.

LÜDELING A, WALTER M., KROYMANN E., ADOLPHS P. (2005). Multi-level error annotation in learner corpora. Actes de *Corpus Linguistics 2005, Birmingham*.

Du TAL dans les écrits scolaires : premières approches

Claire Wolfarth[1] Claude Ponton[1] Catherine Brissaud[1]
(1) Lidilem, CS40700 - 38058 Grenoble cedex 9, France
Claire.Wolfarth@univ-grenoble-alpes.fr, Claude.Ponton@univ-grenoble-alpes.fr, Catherine.Brissaud@univ-grenoble-alpes.fr

RESUME

Dans cet article est présentée une première approche de l'usage de méthodes issues du TAL pour exploiter des textes scolaires, très peu normés. Il permettra d'envisager la spécificité de ces écrits à travers la présentation du corpus étudié avant de se pencher sur les premières hypothèses de traitement automatique en vue d'une annotation des erreurs qui le composent. Y seront également exposés les objectifs de ce travail et la portée attendue.

ABSTRACT

Some NLP in school texts corpora : first hypothesis
In this article, a first approach of non-normed school corpora treatment by using methods from NLP will be exposed. Specific features of this type of texts will be shown by presenting the corpus and first hypothesis to achieve its annotation in term of errors. Main goals of this work will also be presented.

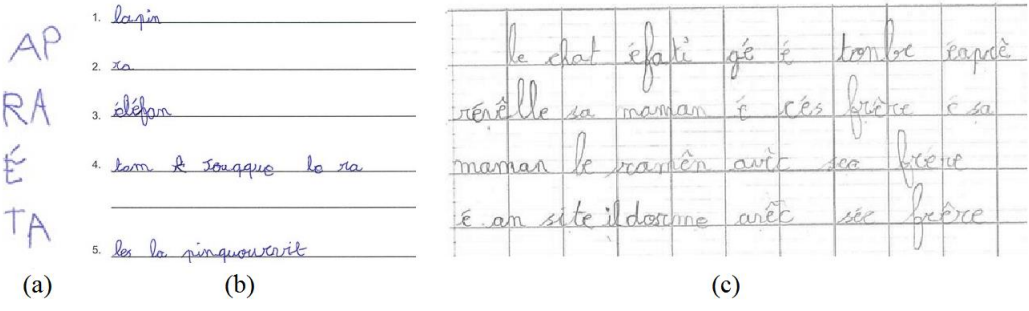
MOTS-CLES : corpus scolaires, annotation automatique, erreurs d'apprenants.
KEYWORDS: school texts corpora, automatic annotation, learner errors

1 Introduction

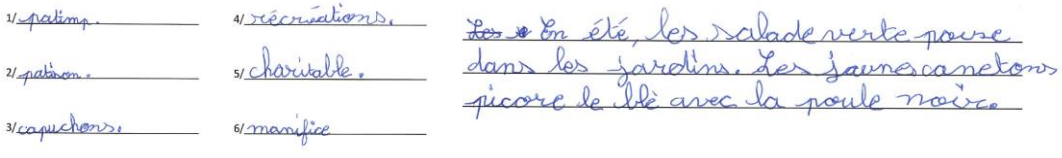
Le travail proposé s'inscrit dans un projet visant à élaborer un large corpus d'écrits d'apprenants accompagné de modules spécifiques d'analyse afin d'outiller la didactique de l'écriture. Ce corpus devrait permettre la description des structures linguistiques utilisées par des élèves en cours de construction de leurs apprentissages de l'écrit à différents niveaux de fonctionnement linguistique (morphographie, syntaxe, lexique, structuration du discours) dans le but de rendre compte des évolutions des procédés d'écriture à différents moments de la scolarisation à l'école primaire. Si certains travaux se sont déjà penchés sur ces questions de description (Elalouf, 2005 ; Auriac-Slusarczyk, Gunnarsson, 2014), ils s'appuient généralement sur des corpus restreints. A terme, la constitution d'un tel corpus et son exploitation devraient permettre de déboucher sur le développement d'activités didactiques adaptées.

Le corpus visé rassemblera différents textes et dictées produits selon un même protocole par les mêmes élèves à des niveaux d'apprentissage différents. Ces productions sont recueillies à chaque fin d'année scolaire du CP au CM2 (2014-2018) pour les textes, en début d'année de CP et en fin d'année de CP et de CE1 pour les dictées. Les élèves concernés sont répartis dans 55 écoles de 5 académies (Grenoble, Clermont-Ferrand, Bordeaux, Lyon et Toulouse soit 1.230 élèves de CP). À terme, ce corpus devrait contenir plus de 3000 productions, ce qui représente, pour le domaine des

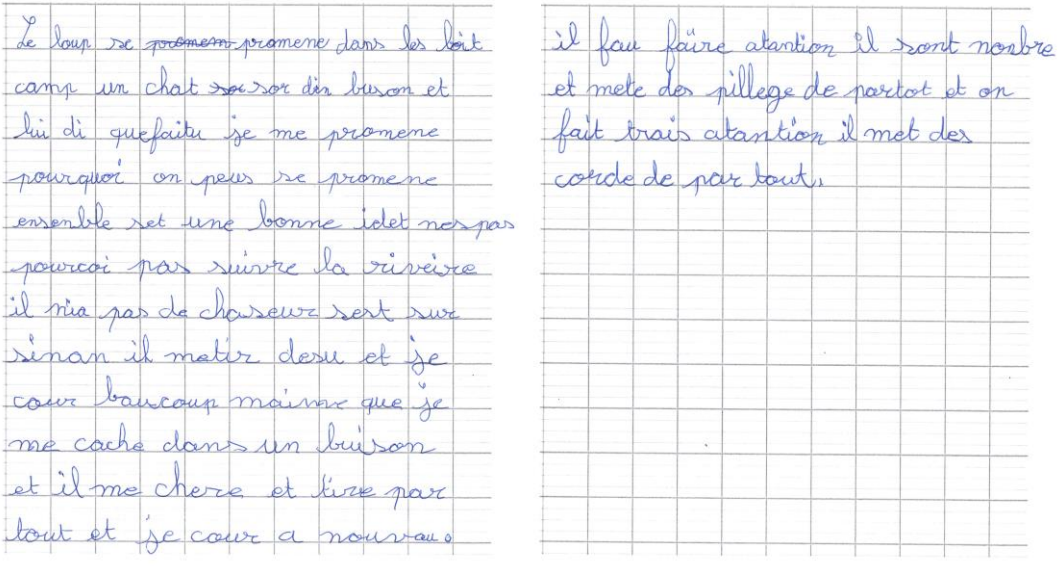
corpus scolaires, un corpus longitudinal de grande taille. Actuellement, seuls les recueils en classe de CP et de CE1 ont été réalisés et seules celles de CP ont été transcrites. Ci-dessous, différents exemples de productions sont donnés.



EXEMPLE 1 : Productions de l'élève 1558 en classe de CP. (a) Dictée produite au mois de septembre. Les mots et phrases dictés sont "lapin", "rat", "éléphant" et "Tom joue avec le rat". (b) Dictée produite au mois de juin. Les mots et phrases dictés sont "lapin", "rat", "éléphant", "Tom joue avec le rat" et "les lapins courent vite". (c) Texte produit au mois de juin. L'enfant devait écrire un texte à partir de 4 images.



EXEMPLE 2 : Dictée produite par l'élève 1558 au mois de juin de la classe de CE1. Les mots et phrases dictés sont "patin", "pâtisson", "capuchon", "récréation", "charitable", "magnifique" et "En été, les salades vertes poussent dans les jardins. Les jeunes canetons picorent le blé avec la poule noire".



EXEMPLE 4 : Texte produit par l'élève 1558 au mois de juin de la classe de CE1. L'enfant devait écrire un texte à partir d'un personnage choisi parmi 4 vignettes.

Afin, de permettre une exploitation riche de ce corpus par les enseignants, les linguistes et les didacticiens, une annotation en terme d'erreurs et de phénomènes notables est prévue. Au vu de la taille du corpus, un recours à des méthodes et outils issus du traitement automatique des langues (TAL) est envisagée. Ces méthodes devraient contribuer à la fois à l'annotation et à l'exploitation du corpus. L'enjeu du projet global est donc triple : 1/ un enjeu linguistique de constitution d'une ressource outillée pour la recherche en linguistique ; 2/ un enjeu pour le TAL, de caractérisation et de modélisation de types d'écrits souvent très éloignés de la norme ; 3/ un enjeu pédagogique et didactique appuyé par la connaissance fine des acquis et difficultés, accessibles au travers d'un outil d'interrogation du corpus. Cet article présente les premières approches exploratoires effectuées dans cette optique sur les productions de CP. Elles constituent le point de départ de différents travaux en cours ou à venir aussi bien en synchronie sur les différents niveaux qu'en diachronie du CP au CM2.

2 État de l'art

Depuis les années 1980, le TAL est étroitement associé à la linguistique de corpus (Habert, Nazarenko, Salem, 1997 ; Kennedy, 1998) par les méthodes et outils qu'il offre pour concevoir et exploiter de grandes masses de données. On trouve dans la littérature autour du TAL un ensemble de travaux connexes au traitement des écrits scolaires, parmi lesquels différentes recherches concernent l'apport du TAL au domaine de l'apprentissage des langues avec notamment les différentes approches du traitement de l'erreur décrites dans l'ouvrage de Heift, Schulze (2007). Pour le français, le projet Freetext (Granger, Vandeventer, Hamel, 2001) mène un travail autour de la

détection automatique d'erreurs basée sur le corpus d'apprenants FRIDA. Ce corpus est également au cœur du projet Exxelant (Antoniadis, Ponton, Zampa, 2010) avec le développement d'un système d'interrogation portant à la fois sur les productions et les corrections.

Même si elle s'adresse le plus souvent à des scripteurs experts, la correction automatique de textes est également un domaine important du TAL proposant des approches variées (Kukich, 1992) potentiellement intéressantes pour le traitement des écrits scolaires. Le traitement automatique d'écrits peu normés est un domaine plus récent en TAL avec notamment les travaux autour de corpus SMS ou issus des réseaux sociaux. Bien que spécifique, ce type de corpus présente des similarités avec les corpus scolaires puisque l'on y retrouve, entre autres, des problématiques liées à la proximité avec l'oral ou à la segmentation en mots (Fairon, Klein, Paumier, 2006). Toujours dans le domaine de ces écrits peu normés, signalons également le système d'étiquetage morphosyntaxique MElt (Denis, Sagot, 2012) développé spécifiquement pour ce type d'écrits et notamment appliqué à des textes provenant de forums en ligne (Baranes, 2012).

3 Spécificités du corpus, hypothèses et premières approches

Outre l'étape de transcription[1] non triviale de ce type d'écrits, son caractère très peu normé constitue un défi pour le TAL. Bien que de nombreux travaux se soient intéressés aux corpus peu normés, il n'est pas possible de transposer ces travaux à notre corpus sans les adapter. En effet, comme le montre les exemples de production précédents, celui-ci présente de nombreuses particularités qui soulèvent des problèmes spécifiques. Pour le moment, seules les productions de CP ont été entièrement transcrites mais nous n'avons pas encore décrit finement, et encore moins quantifié, les phénomènes rencontrés. Toutefois, parmi les erreurs spécifiques, deux grandes catégories semblent se dégager. La première concerne les erreurs de segmentation, témoins d'une connaissance encore floue des frontières de mots chez certains élèves. On donnera pour exemple la production de l'élève 1346 en classe de CP « Le petit chat sanva pan dan cesa maman dore. [...] » (*Le petit chat s'en va pendant que sa maman dort*.), où l'on notera les formes « sanva » et « cesa » dont les formes normées (interprétées) sont respectivement *s'en va* (3 formes) et *que sa* (2 formes) et les formes « pan dan » que l'on peut normaliser par l'unique forme *pendant*. Nous appellerons les premiers exemples des phénomènes d'hyposegmentation et le deuxième un phénomène d'hypersegmentation.

La deuxième catégorie concerne les erreurs de correspondance phonographique pour lesquels l'enfant transcrit le phonème attendu à l'aide d'une graphie non normée (e.g. « tonbe », attendu *tombe*) et d'omission de lettres muettes (e.g. « cha », attendu *chat*). Nos premières constatations, semblent laisser penser que, pour ces erreurs, alors que la norme orthographique n'est pas respectée, les informations phonographiques présentes associées au contexte fournissent des indices forts pour retrouver les formes correctes.

[1] Cette étape consiste en une saisie manuelle, et donc une interprétation, des productions dans une base de données. Plus de détails sur cette opération sont donnés dans (Wolfarth et al., 2016)

À ces deux catégories, il convient de rajouter un ensemble d'erreurs, assez présentes au CP, difficilement interprétables (e.g. élève 56 : « cisr un qui c unqior BouM Miaou q un »).

Au vu de ces spécificités et dans l'état actuel de nos connaissances, il ne nous semble pas envisageable de viser une quelconque automatisation de l'annotation. A l'instar de l'approche moins-disante de Kraif, Ponton (2007), nous situons donc dans une perspective d'aide à l'annotation et à l'exploitation. Nous comptons, d'une part, nous appuyer sur les traitements de plus bas niveau (souvent les plus robustes) pour, ensuite élargir progressivement le spectre de nos analyses. D'autre part, nous faisons l'hypothèse que la connaissance du contexte de production (niveau des scripteurs, consigne de rédaction…) permettra d'affiner la précision des analyses. Le test de cette hypothèse est la base du premier travail que nous avons effectué autour des erreurs à phonologie respectée et que nous présentons brièvement dans le paragraphe suivant.

3.1 Identification des formes normées par comparaison phonologique

En partant du constat qu'un nombre conséquent d'erreurs réalisées en classe de CP modifient la graphie du mot tout en conservant la majeure partie des informations phonologiques, une recherche des formes normées basée sur des comparaisons de formes phonologiques a été élaborée. Pour ce faire, chaque forme a été transcrite sous forme phonologique à l'aide de l'outil LIA_PHON (Béchet, 2001) puis comparée à différentes ressources lexicales plus ou moins proches du contexte de production), elles-mêmes converties sous forme phonologique à l'aide de LIA_PHON. Le résultat de cette comparaison est une liste de formes possibles qu'il reste à désambigüiser.

Parmi les différentes ressources lexicales testées, la ressource Manulex (Manulex CP et Manulex CP-CM2, Ortéga, Lété, 2010) e été retenue, ce choix fait suite à une première étude (Wolfarth et al., 2016) qui semble confirmer l'hypothèse que la connaissance du contexte de production permet d'améliorer sensiblement la qualité du processus d'annotation par le recours à des ressources spécifiques au contexte. En effet, contrairement aux lexiques généraux, le recours à des lexiques proches du vocabulaire utilisé en primaire permet une amélioration nette du processus de désambigüisation.

Cependant, pour un nombre certain de formes toutes les informations phonologiques ne sont pas disponibles, une telle méthode ne permet pas alors de retrouver la forme attendue. Au vu de ces limites et dans notre perspective d'aide à l'annotation et à l'exploitation du corpus, une autre piste autour de l'apport du TAL est à l'étude.

3.2 Annotation par alignement avec un « corrigé »

Si le recours au contexte de production améliore l'analyse, elle est donc loin d'être suffisante. L'idée que voudrions développer dans la suite de nos travaux est de s'appuyer sur une correction des productions pour pouvoir, à travers des comparaisons entre corrections et productions, proposer des analyses fines des erreurs. Si les corrections sont connues pour les activités de dictée, elles seront proposées manuellement lors de l'étape de transcription pour les productions. Cette analyse

comparative nécessitera au préalable une étape d'alignement entre correction et production. Nous comptons pour cela nous appuyer sur les marques phonologiques fortes présentes dans les textes. La comparaison entre les unités alignées se fera sur différents niveaux (graphiques, lexicaux, morphologiques, phonétiques, etc.) comme proposé dans le travail de Blanchard et al. (2009).

Un premier travail selon cette méthode est en cours sur les dictées réalisées au mois de juin en classe de CP et plus spécifiquement sur les deux phrases dictées ("Tom joue avec le rat" et "les lapins courent vite"). Dans ce cadre, un premier aligneur dictées/corrections ainsi qu'un premier module de comparaison sont en cours de développement. L'aligneur s'appuie sur les formes phonologiques et permet de produire des appariements entre segments comme [*courvit*] avec [*courent vite*]. La première version de l'analyseur de différence s'intéresse à la détection des cas d'hyper et d'hypo-segmentation, aux ajouts/omissions de mots, au niveau graphique (casse, différence de caractères) et au niveau phonologique (respect ou non de la phonologie). Un travail aux niveaux des graphèmes, des syllabes et de la morphologie est prévu.

4 Perspectives

Cette étude constitue une première approche de l'usage de méthodes issues du TAL sur des textes scolaires très peu normés. Les deux approches décrites précédemment sont en cours d'amélioration et d'évaluation. De ces différentes études sont attendus différents apports pour l'étude de l'apprentissage de l'écriture. En effet, il s'agira de mettre à disposition des communautés scientifiques et pédagogiques un corpus outillé, riche et unique pour sa taille et, surtout, pour son aspect longitudinal. D'autre part, le développement de méthodes et d'outils TAL spécifiques pour sa constitution et son interrogation permettra une approche fine de l'ensemble des données qui resterait difficile de manière manuelle.

Sur le plan linguistique, une telle ressource devrait permettre de mieux connaitre les phénomènes à l'œuvre dans le processus d'acquisition de l'écriture. D'un point de vue didactique, elle devrait appuyer la réflexion pédagogique des enseignants par des exemples réels de productions montrant ainsi les connaissances et les difficultés à chaque niveau. À terme, les développements TAL mis au point lors de la constitution du corpus devraient permettre, par exemple, le développement d'activités didactiques autour des difficultés repérées en corpus.

Dans le cadre global de ce projet, il nous semble nécessaire de maintenir des liens forts entre la recherche et le terrain. La constitution de ce corpus longitudinal sur cinq années nécessite un réseau d'enseignants impliqués et motivés. Dans cette optique, nous mettons à disposition de ce réseau, les corpus transcrits au fur et à mesure de leur conception sous forme de site web associé à certaines fonctions d'exploitation. Le site sur le corpus de CP sera ouvert dans les jours à venir et nous comptons sur les retours de ces utilisateurs pour affiner nos outils avant une mise à disposition globale auprès des enseignants et des chercheurs à la fin du projet.

Références

ANTONIADIS G., PONTON C., ZAMPA V. (2010). Exxelant et Mirto – Deux exemples d'environnement d'ALAO intégrant des outils TAL. *Multilinguisme et traitement des langues naturelles*. Montréal, Canada : PUQ.

AURIAC-SLUSARCZYK E., GUNNARSON-LARGY C. (2014). *Écriture et réécritures chez les élèves: un seul corpus, divers genres discursifs et méthodologies d'analyse*. Academia.

BARANES M. (2012). Vers la correction automatique de textes bruités: Architecture générale et détermination de la langue d'un mot inconnu. Actes de *RECITAL'2012 - Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, 95-108.

BECHET F. (2001). LIA_PHON - Un système complet de phonétisation de textes, *Traitement Automatique des Langues (T.A.L.)* 42(1), 47-67.

BLANCHARD A., KRAIF O., PONTON C. (2009). "Mastering Overdetection and Underdetection in Learner-Answer Processing: Simple Techniques for Analysis and Diagnosis". *Calico Journal*, *Vol. 26*(No. 3), 592-610.

DENIS P., SAGOT B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation* 4(46), 721-736.

ELALOUF M.-L. (dir) (2005). *Écrire entre 10 et 14 ans. Un corpus, des analyses, des repères pour la formation*, SCérén, CRDP de Versailles.

FAIRON C., KLEIN J.R., PAUMIER S. (2006). *Le langage sms. Étude d'un corpus informatisé à partir de l'enquête « Faites don de vos sms à la science »*. Louvain-la-Neuve : Presses universitaires de Louvain.

GRANGER S., VANDEVENTER A., HAMEL M.-J. (2001). Analyse des corpus d'apprenants pour l'ELAO basée sur le TAL. *Traitement automatique des langues* 42(2), 609-621.

HABERT B., NAZARENKO A., SALEM A. (1997). *Les linguistiques de corpus*. Paris : Colin.

HEIFT T., SCHULZE M. (2007). *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*, New York and London : Routledge.

KENNEDY G. (1998). *An introducing to Corpus Linguistics*. London and New York : Longman.

KRAIF O., PONTON C. (2007). Du bruit, du silence et des ambiguïtés: que faire du TAL pour l'apprentissage des langues. Actes de *TALN* (Toulouse).

KUKICH K. (1992). *Techniques for Automatically Correcting Words in Text*. ACM Computing Surveys 24 (4), 377–439.

ORTÉGA É., LÉTÉ B. (2010). « eManulex: Electronic version of Manulex and Manulex-infra databases », http://www.manulex.org

Wolfarth, C., Ponton, C., Totereau, C. (2016). Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire. Corpus.

Élaboration semi-automatique d'une ressource de patrons verbaux

Sylvain Hatier[1] Rui Yan[1]
(1) LIDILEM – EA 609 – Université Grenoble Alpes
sylvain.hatier@univ-grenoble-alpes.fr, rui.yan@univ-grenoble-alpes.fr

RÉSUMÉ

Nous présentons dans cet article l'élaboration semi-automatique d'une ressource de patrons verbaux. Nous nous intéressons aux verbes du Lexique Scientifique Transdisciplinaire, lexique essentiel dans l'argumentation et l'organisation du discours dans les écrits scientifiques. Ce travail se base sur un corpus arboré d'articles scientifiques à partir duquel sont extraits les cadres de sous-catégorisation. Ces cadres sont alors manuellement regroupés en patrons, sur le modèle CPA, puis associés à une acception définie pour chaque verbe. La ressource résultante, élaborée dans un but d'aide à la rédaction et à la compréhension scientifique, répertorie les constructions verbales les plus fréquentes dans ce genre. L'utilisateur a ainsi accès, pour chaque acception verbale, aux cooccurrents et constructions syntaxiques préférentiels.

ABSTRACT

Semi-automatic elaboration of a verbal patterns dictionary

This study focuses on the semi-automatic elaboration of verbal patterns resource. We explore more specifically Cross-Disciplinary Lexicon verbs, a lexicon which plays an essential role in reasoning and discourse organizing in academic writing. This work is based upon a parsed corpora of French scientific articles, from which we automatically extracted subcategorization frames. These frames are then manually gathered, according to CPA model, and combined with a verbal meaning. The verbal patterns resource, designed for the didactic purpose of academic writing and reading, lists the most frequent verbal constructions in this genre. User thus has access, for each verbal meaning, to preferential syntactical collocates and constructions.

MOTS-CLES : écrit scientifique, sous-catégorisation verbale, linguistique de corpus, patron lexico-syntaxique.

KEYWORDS : academic writing, verbal subcategorization, corpus linguistics, lexico-syntactic pattern.

1 Cadres et objectifs

Ce travail d'élaboration d'une ressource lexicale dédiée aux constructions verbales dans l'écrit scientifique répond à plusieurs observations. Nous avons constaté dans de précédents travaux (Hatier, Yan, 2015) la difficulté des étudiants à maîtriser les patrons verbaux dans l'écrit scientifique, à la suite des travaux de Nesselhauf (2005), Hyland (2008) ou Granger, Paquot (2009). Ces difficultés sont d'ailleurs partagées par les apprenants et les étudiants francophones natifs (Hatier, Yan, à paraître). Notre objectif, d'ordre didactique, est ainsi de proposer une ressource des

patrons verbaux permettant de remédier à ces lacunes, afin d'améliorer la rédaction et la compréhension de textes scientifiques. La ressource est destinée aux enseignants, afin de le permettre la mise en place de séquences didactiques adaptées au genre de l'écrit scientifique. Nous adoptons pour cela une approche de linguistique de corpus, en faisant émerger du corpus d'analyse, représentatif du genre étudié, les constructions correspondant à l'usage.

Nous nous intéressons plus spécifiquement aux verbes du lexique scientifique transdisciplinaire (LST). Ce lexique méta-discursif et méta-scientifique est associé au genre de l'écrit scientifique. Nous le définissons, à la suite de Tutin (2007), comme un lexique renvoyant au discours sur les objets et les procédures scientifiques. Le LST (*hypothèse, analyser, qualifier,* etc.) est un lexique essentiel à maîtriser dans l'argumentation, l'organisation textuelle et l'expression de l'opinion.

1.1 Lexique Scientifique Transdisciplinaire

Dans une étude précédente (Hatier et al., 2014), nous avons procédé à l'extraction d'une liste de 1312 mots du LST : 274 adjectifs, 202 adverbes, 493 noms et 342 verbes. L'ensemble a ensuite été organisé en une classification sémantique en classes et sous-classes transcatégorielles, en se basant sur les propriétés sémantiques et lexico-syntaxiques des unités lexicales. Ces classes sont des ensembles de co-hyponymes, mots du LST, confirmant une définition et un test lexico-syntaxique définitoire de la classe. Ainsi, la sous-classe {document} de la classe {communication}, ayant notamment pour membre *article, ouvrage, texte*, a pour test d'appartenance : *Ce N présente*. En disposant d'une telle classification, nous pouvons intégrer dans l'analyse des constructions verbales le niveau sémantique, au niveau du verbe analysé et de ses arguments. De plus, cette typologie nous permet d'étudier les patrons verbaux au sein des classes sémantiques homogènes et de proposer une entrée onomasiologique dans notre ressource adaptée pour l'aide à la rédaction scientifique. L'identification des acceptions verbales du LST a été effectuée en prenant comme référence la ressource *Les Verbes Français*[1] (LVF, cf Dubois, Dubois-Charlier, 1997).

1.2 Patrons verbaux dans l'écrit scientifique

Nous nous intéressons à l'étude des patrons lexico-syntaxiques des verbes du LST, selon une approche contextualiste (Sinclair, 1991 ; Hunston, Francis, 2000 ; Hanks, 2008). Un patron consiste en « une structure syntaxique intégrant des collocations privilégiées » (Hanks, 2013 : 92, traduit par l'auteur). Il est défini dans le cadre du modèle Corpus Pattern Analysis (CPA) de Hanks (*ibid.*) et se caractérise par 1) l'association entre le sens et l'usage réel du mot 2) l'étiquetage sémantique au niveau des arguments. Un des patrons du verbe *to execute* est ainsi représenté comme suit :

> Pattern[2]: [[Human | Institution]] execute [[Plan | Command | Activity]]
> Implicature : [[Human | Institution]] does work in order to put [[Plan | Command | Activity]] into effect.
> Ex : The **student** has the opportunity to formulate and **execute a search**. (L'étudiant a la possibilité de concevoir et d'exécuter une recherche)

Dans ce patron, le verbe *to execute* sélectionne de préférence des sujets noms renvoyant à un humain / une institution et des objets noms renvoyant à un plan / un ordre / une activité. Le sens

1 Consultable en ligne : http://rali.iro.umontreal.ca/rali/? q=fr/node/1237/ [consulté le 25/04/2016]
2 http://www.pdev.org.uk/#browse ? q=; f=C [consulté le 25/04/2016]

associé à ce patron est indiqué dans l'*implicature*, au travers d'une paraphrase. Le patron permet ainsi de faire le lien entre sens mobilisé et constructions verbales. Les verbes étant polysémiques, y compris dans le contexte de l'écrit scientifique, l'ambiguïté peut être ainsi levée en associant les valeurs sémantiques des arguments aux structures syntaxiques. Le modèle CPA accorde un rôle primordial aux contextes et aux usages en corpus, ce qui nous paraît adapté à notre perspective didactique.

2 Méthodologie

L'élaboration de la ressource de patrons verbaux s'effectue en trois phases principales. Dans un premier temps, nous constituons et annotons le corpus d'analyse. De ce corpus sont ensuite extraits automatiquement les cadres de sous-catégorisation des verbes du LST. Ces cadres sont alors regroupés sous forme de patrons lexico-syntaxiques modélisés manuellement.

2.1 Corpus d'analyse

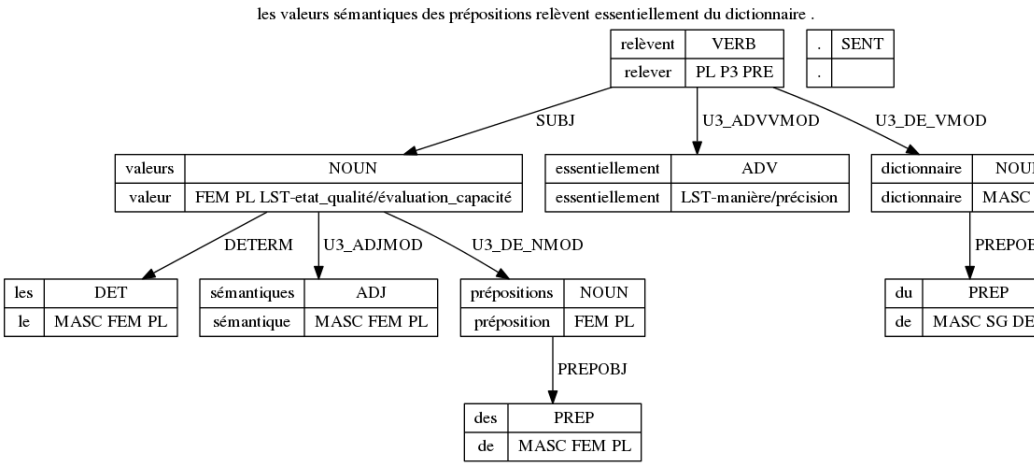
Le corpus d'analyse, de 5 millions de mots, issu du projet scientext, est composé de 500 articles de recherche en français (Tran, 2014), de 10 disciplines des sciences humaines et sociales. À l'aide de l'analyseur en dépendances XIP[3] (Aït-mokhtar et al., 2002), il a ensuite été annoté en lemmes, traits morpho-syntaxiques et relations syntaxiques. Les traits de classes et sous-classes sémantiques ont ensuite été projetés dans le corpus afin d'intégrer ces informations sémantiques dans les patrons verbaux extraits. Nous avons également procédé à un post-traitement du corpus en définissant des grammaires locales en vue de l'amélioration de l'annotation en dépendances. Nous souhaitons ainsi améliorer l'analyse à travers notamment la propagation du sujet pour les structures avec verbes de contrôle, ou lorsque le sujet est en situation de coréférence[4] ou de coordination.

2.2 Extraction des cadres de sous-catégorisation

À l'instar de Messiant et al. (2010), nous procédons alors à l'extraction des cadres de sous-catégorisation « sans a priori, pour faire émerger du corpus [ceux] correspondant à l'usage ». Nous commençons par extraire autant de cadres que d'occurrences pour chaque verbe. Les cadres sont ensuite regroupés, lors d'une phase de factorisation, à la manière de Kupsc (2007). Certaines relations de dépendance sont rassemblées sous une même étiquette (telles les relations avec un semi-modal : *devoir, sembler, aller*). Chaque occurrence d'un verbe du LST est ainsi représentée en termes d'ensemble de relations (impliquant le verbe en tant que gouverneur ou dépendant) pour aboutir aux cadres de sous-catégorisation, tel que l'illustre la figure ci-après :

3 Nous remercions Claude Roux pour ces précieux conseils dans le paramétrage et l'utilisation de XIP

4 Considérons la phrase « *Nous analysons la préposition qui introduit le nom* ». Si l'analyseur syntaxique définit une relation de coréférence entre *qui* et *préposition* et une relation sujet entre *introduit* et *qui*, alors nous propageons la relation sujet entre *introduit* et *préposition*.



Étant donné que nous ne retenons pas les relations avec les adverbes, le verbe *relever* est, dans cet exemple, impliqué dans deux relations :

- gouverneur dans la relation “sujet” avec pour dépendant *valeurs* (nom du LST, dont la classe et la sous-classe sont renseignées dans les traits) ;
- gouverneur dans la relation 'complément prépositionnel' avec pour dépendant *dictionnaire.*

La phrase illustrée ci-dessus a alors pour cadre de sous-catégorisation :

(SUBJ {document}) relever_VERB (COMPLEMENT PREP.).

Le verbe *relever*, dans cette construction transitive indirecte, prend pour sujet un nom du LST de la classe sémantique {document] et pour complément un syntagme nominal introduit par une préposition.

Pour chaque cadre, sont indiqués les lemmes et les classes sémantiques les plus fréquents en tant qu’argument. Le tableau suivant présente deux des cadres de sous-catégorisation les plus fréquents dans notre corpus.

Cadre – *Exemple*	Fréquence	Sujet	Objet
(Subj – hum) montrer (Complétive QUE) *Les résultats montrent que les effets sont significatifs*	762	résultat, il, analyse, étude, travai	-
(Subj – hum) constituer (Obj – hum) *Ils constituaient un groupe solidaire et homogène*	616	il, pratique, type, celui-ci, espace	élément, enjeu, point, forme

TABLE 1: Exemples de cadres de sous-catégorisation

Ces cadres résultent d'un premier regroupement, opéré automatiquement, d'occurrences d'un même verbe et permettent le passage à la phase de modélisation manuelle des patrons, étape essentielle pour le traitement de la polysémie.

2.3 Modélisation des patrons verbaux

Afin de modéliser les patrons verbaux, nous nous basons dans un premier temps sur les cadres de sous-catégorisations extraits pour repérer les acceptions verbales spécifiques à l'écrit scientifique, utilisées dans l'argumentation et la présentation de l'activité scientifique. Nous mettons ainsi en correspondance un cadre avec un sens défini, présent dans notre classification du LST. Dans un second temps, l'observation du corpus et l'analyse des cadres de sous-catégorisations nous permettent d'aboutir aux patrons verbaux du LST.

Nous détaillons dans la section suivante ces étapes d'analyse et, en partant d'exemples issus de notre corpus, illustrons la modélisation des patrons.

2.3.1 Repérage des acceptions

Le repérage des acceptions constitue l'étape de base de notre analyse étant donné la polysémie des verbes. Prenons par exemple le verbe *postuler,* dont un des cadres extraits est : (SUBJ +hum) postuler_VERB(COMPLETIVE_QUE)

Ce cadre s'actualise dans l'exemple ci-dessous :

Ex : *Dans cet article,* ***nous postulons que*** *l'observation des phénomènes non-verbaux telle que nous la concevons est susceptible d'enrichir une approche didactique.*

Le sens mobilisé correspond à l'entrée 2 du verbe *postuler* dans le LVF, dont la définition est « supposer » (ex. : *On postule la bonne foi des joueurs/que l'accord est possible*.). Cette acception apparaît dans une construction avec un sujet humain et soit un objet chose soit une complétive.

2.3.2 Exemple de patrons lexico-syntaxiques

Après avoir relié cadres de sous-catégorisation et acception, nous procédons à la modélisation des patrons verbaux. Nous détaillons dans cette partie des exemples de patrons pour des verbes renvoyant à une ou plusieurs acceptions dans le corpus d'analyse.

Lorsqu'un verbe est monosémique dans le corpus, ce sens peut correspondre à plusieurs patrons. Par exemple, le verbe *constater*, appartenant à la sous-classe {analyse_info/constat}, prend pour sujet un nom humain et comme objet soit un complément nominal soit une complétive conjonctive. Ici, il prend le sens « remarquer » (entrée *constater 01* dans le LVF). Cette acception est réalisée à travers les deux patrons suivants :

1. [[humain=auteur, chercheur]] constate [complétive *que*]
   Ex : ***On constate que*** *le taux d'actualisation optimal est en général au-dessus* […].

2. [[humain=auteur, chercheur]] constate [[phénomène|relation]]
   Lexset [[phénomène|relation]] : *différence, effet, écart, inégalité*
   Ex : *Deuxièmement,* ***nous constatons trois problèmes*** *qui concernent uniquement* […].

Pour chaque cadre sont renseignées les informations sur les cooccurrences statistiquement significatives au niveau des actants. Ceci nous permet d'indiquer le(s) type(s) sémantique(s) approprié(s). Notons qu'il ne s'agit pas de représenter l'ensemble des cooccurrents possibles, mais de présenter un usage prototypique permettant de clarifier l'acception.

Dans les cas où un verbe renvoie à plusieurs acceptions, nous vérifions si le type de construction syntaxique permet de désambiguïser. Ainsi, le verbe *considérer* possède trois sens correspondant à trois constructions différentes :

1. [[humain=auteur, chercheur]] (peut) considérer [[entité abstraite|éventualité]] comme
   SENS : Quelqu'un donne son jugement par rapport à quelque chose
   EX : *On peut dire la même chose des niveaux d'action, c'est-à-dire des niveaux que* ***l'on considère comme*** *pertinents* […]

2. [[humain=auteur, chercheur]] (peut) considérer [complétive que]
   SENS : Quelqu'un se forme une opinion autour d'un fait scientifique
   EX : *Sur le plan symbolique,* ***on pourrait considérer que*** *la figure de professeur renvoie à une forme hiérarchique plus explicite* […]

3. Si/Lorsque [[humain=auteur, chercheur]] considère [[concept|éventualité]]
   SENS : Si/Lorsque quelqu'un donne son attention à quelque chose pour un examen attentif ou critique, introduisant ainsi un nouveau thème
   EX : ***Si l'on considère*** *par exemple la notion de passé,* […]

La polysémie peut se situer au niveau des arguments, l'attribution des types sémantiques joue ainsi un rôle important pour distinguer les différents sens. Par exemple, le verbe *caractériser* apparaît dans la construction transitive directe. Au niveau du sujet, nous pouvons distinguer deux types sémantiques distincts : [[humain]] (*nous, on,* etc.) et [[entité abstraite]] (*propriété, élément,* etc.). Le premier type est associé au sens « indiquer le trait de », tandis que le second active le sens « constituer le trait de », comme illustré dans les deux patrons suivants :

1. [[humain]] caractérise [[évènement|entité abstraite]]
   Ex : *Dans la proposition suivante,* ***nous caractérisons la probabilité*** *de contrôle optimale.*

2. [[entité abstraite 1=trait]] caractérise [[évènement|entité abstraite 2]]
   Ex : ***L'indécidabilité*** *qui* ***caractérise ces faits*** *finit alors par se conjuguer […].*

L'analyse des patrons permet ainsi de représenter les propriétés syntaxiques et sémantiques du verbe, dans un usage spécifique à l'écrit scientifique, assurant l'adaptation de la ressource de patrons à l'aide à la rédaction scientifique.

Au niveau de la ressource des patrons, une entrée correspond à une acception verbale qui peut être mobilisée par un ou plusieurs patrons, comme l'illustre l'entrée d'une acception du verbe *montrer* ci-dessous :

[[objet scientifique]] (sembler) montre [complétive que ou comment]
SENS : Quelque chose révèle ou atteste qu'un fait scientifique est juste et le met en évidence, ceci constitue un argument important de l'auteur.
LEXSET : [[objet scientifique]] : résultat, exemple, tableau, enquête, donnée, entretien
COMMENTAIRE : le gérondif est aussi fréquent

ROUTINES : *ce résultat tend à montrer que, cela montre que*
EX : *Le tableau V montre par ailleurs que l'apprentissage du solfège profite aux enfants issus des classes sociales favorisées quelle qu'en soit la durée alors que seul un apprentissage durable (d'au moins deux ans) a un effet sur les résultats des élèves issus des classes sociales non favorisées*. (scienceseducation.xml-s1945)

3 Conclusion

La combinaison de traitement automatique sur les cadres de sous-catégorisation et de modélisation manuelle de patrons lexico-syntaxiques nous a permis d'élaborer une ressource lexicale des patrons verbaux dans l'écrit scientifique. L'intégration de traits sémantiques au niveau des arguments permet de faciliter l'identification des sens mobilisés et des emplois adaptés. Nous avons pour perspective l'amélioration de la phase de factorisation afin d'alléger la tâche d'observation des cadres qui précède la modélisation des patrons. Ainsi, bien que ces regroupements réduisent de moitié les cadres à analyser (le verbe *considérer,* 2559 occurrences, passe de 1119 cadres à 569, le verbe *montrer,* 3114 occurrences, passe de 1137 cadres à 522), le nombre de cadre correspondant à une unique occurrence reste trop élevé (569 pour *considérer* contre 864 avant factorisation, 355 pour *montrer* contre 857 avant factorisation). Une autre amélioration se situe au niveau de l'identification des erreurs d'analyse syntaxique qui donnent lieu à des cadres dont un ou plusieurs arguments sont manquants. Ainsi, dans la phrase suivante, le verbe *correspondre* n'entre dans aucune relation selon l'annotation syntaxique : *à chaque niveau de segmentation institutionnalisé correspond un doyen*. Il résulte de ce genre d'erreur du bruit dans la liste de cadre (en intégrant un cadre tronqué) et du silence (en ne faisant pas correspondre ces occurrences avec le bon cadre).

En termes d'utilisation de la ressource, nous en distinguons deux principales. La première, d'ordre didactique, se situe dans le cadre de l'aide à la rédaction. La ressource peut faciliter alors l'encodage et/ou le décodage de constructions verbales typiques de l'écrit scientifique. Ainsi, au niveau du décodage, l'enseignant peut proposer à l'apprenant de repérer les différentes acceptions du verbe *considérer* pour un ensemble d'exemples sélectionnés (voir le tableau 2). L'enseignant choisit des exemples correspondant aux patrons sur lesquels il veut faire travailler les apprenants, pour que ceux-ci identifient les différentes acceptions mobilisées selon la construction employée.

Cela est vrai, également, si	l'on considère	l'immense importance qu'ont prise les ONG dans la vie du monde…	sociologie.xml-s10035
Patrick Charaudeau propose de	considérer	tout fait humoristique comme un acte de langage.	scinfo.xml-s10057
En effet, leurs pratiques d'information se sont révélées très disparates, certains de ces jeunes	considérant	le Net bien plus comme un moyen de communication et de jeu...	scinfo.xml-s10171
En ce qui concerne la dénomination des types de mobilité, nous	considérons	qu'il existe trois critères de distinction : la dimension spatiale (mobilité locale / régionale / nationale / internationale)… [4].	geo.xml-s11184
Premièrement, lorsqu'on	considère	le type d'interventions mises en place après le diagnostic, on constate qu'il s'agit principalement d'interventions centrées sur les conditions de travail.	psycho.xml-s592

TABLE 2: Concordancier du verbe *considérer*

À l'issue de ce type de séquence didactique, l'apprenant peut être amené à réinvestir ces connaissances en employant à son tour les patrons dans des exercices de rédaction scientifique.

La seconde utilisation que nous envisageons permettrait d'interroger un corpus analysé en dépendances afin d'identifier automatiquement la réalisation d'un patron (et de son acception correspondante) en fonction des éléments le constituant (un lemme verbal, un ensemble de relations syntaxiques, des éventuels traits sémantiques pour les arguments).

Remerciements

Nous souhaitons remercier la région Rhône-Alpes pour le financement de nos travaux de recherche ainsi que les partenaires du projet ANR-Contint Termith [5] pour leur collaboration dans nos expérimentations sur le LST.

Références

AÏT-MOKHTAR S., CHANOD J.-P., ROUX C. (2002). Robustness beyond shallowness : incremental deep parsing. *Natural Language Engineering*, *8* (2-3), 121–144.

DUBOIS, J., DUBOIS-CHARLIER, F. (1997) : *Les verbes français*. Larousse.

GRANGER S., PAQUOT M. (2009). Lexical Verbs in Academic Discourse : A Corpus-driven Study of Learner Use. In : Charles, Maggie/Hunston, Susan/Pecorari, Diane (éds.) : *Academic Writing : At the Interface of Corpus and Discourse*. London : Continuum International Publishing Group :193-214.

HANKS P., PUSTEJOVSKY J. (2005). A Pattern Dictionary for Natrual Language Processing. *Revue française de Langue Appliquée*10 (2), 63-82.

HANKS P. (2008). Lexical Patterns : from Hornby to Hunston and beyond. In E. Bernal et J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress.* Barcelona : IULA : 89-129.

HANKS P. (2013). *Lexical Analysis : Norms and Exploitations.* MIT Press.

HATIER S., TUTIN A., JACQUES M.-P., JACQUEY E., KISTER L. (2014). Catégorisation sémantique des noms simples du lexique scientifique transdisciplinaire. Présentation à *ACFAS, colloque 330 : Étude de lexiques à vocation particulière : approches théoriques, méthodologiques, pédagogiques et multidisciplinaires*, Montréal.

HATIER S., YAN R. (2015). Comparaison de constructions verbales entre un corpus d'apprenants et un corpus d'articles de recherche. Présentation à *8es Journées Internationales de Linguistique de Corpus (JLC2015)*, Orléans.

HATIER S., YAN R. (à paraître). Analyse contrastive des patrons verbaux dans l'écrit scientifique entre scripteurs étudiants et experts.

HUNSTON, S., FRANCIS, G. (2000). *Pattern Grammar : a corpus-driven approach to the lexcial grammar of English*. Amsterdam/Philaelphia : John Benjamins.

5 TermITH (Terminologie et Indexation de Textes en sciences Humaines) : ANR-12-CORD-0029 CONTINT. ATILF, INIST, LIDILEM, LINA, INRIA NGE et Saclay. http://www.atilf.fr/ressources/termith/ (visité le 25 avril 2016)

HYLAND K. (2008). Academic clusters : text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, 18, (1) : 41-62.

JACQUES M.-P. (2011). Nous appelons X cet Y : X est-il un terme émergent ? In K. Kageura & P. Zweigenbaum (Éd.), *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence* (p. 31–37). Paris, France : INALCO.

KUPSC A. (2007). Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré. Présenté à *TALN* 2007.

MESSIANT C., GABOR K., POIBEAU T. (2010). Acquisition de connaissances lexicales à partir de corpus : la sous-catégorisation verbale en français. *Traitement automatique des langues*, *51*(1), 65–96.

NESSELHAUF N. (éd.) (2005) : *Collocation in a Learner Corpus*. Amsterdam / Philadelphia : John Benjamins Publishing Company.

PAQUOT M. (2010) : *Academic vocabulary in learner writing : From extraction to analysis*. Bloomsbury Publishing.

SINCLAIR J. (1991). *LexicalCorpus, concordance, collocation (Vol.1). : Norms and Exploitations*. Oxford University Press Oxford.

TRAN T. T. H. (2014). *Description de la phraséologie transdisciplinaire scientifique et réflexions didactiques pour l'enseignement à des étudiants non-natifs. Application aux marqueurs discursifs* (Thèse de doctorat). Université de Grenoble.

TUTIN A. (2007). Autour du lexique et de la phraséologie des écrits scientifiques. *Revue française de linguistique appliquée*, Vol. XII(2), 5-5.

Exploitation d'une base lexicale dans le cadre de la conception de l'ENPA Innovalangues

Mathieu Mangeot[1,4] Valérie Bellynck[1] Emmanuelle Eggers[4]
Mathieu Loiseau[2,4] Yoann Goudin[3,4]
(1) LIG, Bâtiment IMAG, 700 avenue Centrale - Domaine Universitaire, 38400 Saint-Martin-d'Hères, France
(2) LIDILEM, Université Grenoble Alpes, CS40700, 38058 Grenoble cedex 9
(3) CERLOM, INALCO, 2 rue de Lille, 75007 Paris
(4) IDEFI Innovalangues, Maison des Langues et des Cultures, 1141 avenue Centrale - Domaine Universitaire, 38400 Saint-Martin-d'Hères, France
valerie.bellynck@imag.fr, emmanuelle.eggers@univ-grenoble-alpes.fr, mathieu.loiseau@univ-grenoble-alpes.fr, mathieu.mangeot@imag.fr, yoann.goudin@univ-grenoble-alpes.fr

RÉSUMÉ

Le projet Innovalangues a pour but la conception d'un environnement numérique personnalisé d'apprentissage des langues (ENPA). Dans ce cadre, plusieurs modules tels des jeux sérieux ou des générateurs d'exercices ont besoin d'une base lexicale. Il est intéressant également de donner la possibilité aux apprenants de gérer leurs propres lexiques. Pour éviter que chacun ne développe sa propre base lexicale, il apparaît rapidement indispensable de développer une seule base lexicale commune qui puisse servir aux modules (machines) comme aux apprenants (humains). Après une analyse des besoins autour de scénarios d'utilisation, nous proposons une architecture de base lexicale multilingue. Nous avons ensuite réalisé un prototype fonctionnel qui s'intègre à l'ENPA et permet à un apprenant de consulter des ressources lexicales existantes et de créer son propre lexique. Le prototype LexInnova utilise la plate-forme Jibiki de gestion de ressources lexicales hétérogènes à distance via son interface de programmation (API) REST.

ABSTRACT

Operating a lexical database in the framework of the Innovalangues language learning platform

The Innovalangues project aims to design a personalized digital environment for language learning. In this context, several modules such as serious games or exercise generators need a lexical database. It is also interesting to provide solutions for learners to manage their own lexicons. To prevent each one to develop its own lexical database, it quickly appears essential to develop one common lexical database that can be used for modules (machines) and learners (humans). After an analysis around usage scenarios, we set up a multilingual lexical database architecture. We then built a working prototype that integrates with the environment and allows a learner to look up existing lexical resources and create its own lexicon. The LexInnova prototype uses the Jibiki platform for managing heterogeneous lexical resources via its REST application programming interface.

MOTS-CLÉS : Innovalangues, base lexicale, dictionnaire, lexique, Jibiki, LexInnova
KEYWORDS : Innovalangues, lexical database, dictionary, lexicon, Jibiki, LexInnova

1 Introduction

L'un des principaux objectifs du projet IDEFI Innovalangues est de mettre à disposition un environnement numérique personnalisé d'apprentissage des langues (ENPA). Outre les parcours d'apprentissage, l'écosystème numérique proposé par Innovalangues fournira aux apprenants des outils pour soutenir l'apprentissage : des jeux, des générateurs d'exercices, un chat, un test de positionnement (test automatique adaptatif pour l'orientation de l'apprenant vers un niveau cible en fin de formation), etc.

Le projet est divisé en plusieurs lots (donnant lieu à des livrables du projet) dont l'objectif est de produire des outils répondant à des problématiques didactiques et liés ou intégrés à l'ENPA.

L'une des tâches fondamentales de l'apprentissage des langues est de travailler le lexique[1]. Dès lors, il est rapidement apparu que la réalisation des divers lots (outils) conduisait au développement de d'autant de bases lexicales avec des contraintes spécifiques pour chacune, sans qu'il n'y ait de lien pratique entre eux. De plus, les apprenants peuvent souhaiter également créer leurs propres lexiques d'apprentissage. Cette situation a rapidement posé un problème de mise en œuvre informatique (opérationnalisation), ce qui a mené à la définition d'un sous projet transversal au projet Innovalangues (« chantier interlot »), que nous avons appelé LexInnova.

Nous présentons d'abord le projet Innovalangues et les besoins les plus emblématiques en termes de ressources lexicales. La partie suivante concerne le cahier des charges de l'outil que nous souhaitons construire ainsi que les différents scénarios d'utilisation des ressources lexicales. Ensuite, nous spécifions la base lexicale multilingue visée. Enfin, dans la dernière partie, nous détaillons le fonctionnement du prototype fonctionnel que nous avons implémenté afin de montrer les possibilités qu'offre l'intégration un tel outil.

2 Présentation du projet Innovalangues

Le projet IDEFI Innovalangues, initié en 2012, est piloté par le service LANSAD de l'Université Grenoble-Alpes. Il tire son origine des difficultés des étudiants à atteindre un niveau B2 du Cadre Européen Commun de Référence pour les Langues (CECRL) (CECRL, 2000) alors qu'il s'agit du niveau cible pour le baccalauréat[2] (Masperi et Quintin, 2014a : 62) . À partir de ce constat, le projet a pour but de fournir aux établissements supérieurs des moyens (dispositifs, méthodes, contenus) pour porter le degré de maîtrise en langues des étudiants à un niveau B2 certifié. Pour y parvenir, Innovalangues s'est fixé trois objectifs principaux :

- « se doter d'un environnement (numérique) personnalisé d'apprentissage (ENPA), “open source”, [proposant des contenus sous licence libre], au service de la collaboration, particulièrement flexible et hautement adaptable aux différents contextes rencontrés sur le terrain de la formation » (Masperi et Quintin, 2014b : 8) ;
- proposer un espace dédié aux enseignants et ainsi « capitaliser les résultats de la recherche en didactique des langues » (Masperi et Quintin, 2014a : 72) par des actions de formation et / ou via l'outillage techno-pédagogique ;
- créer une communauté de pratiques pour « étendre la dynamique de recherche-action en langues entreprise au niveau local » (Masperi et Quintin, 2014a : 72).

Le projet se déroule principalement en deux phases. La première phase est orientée conception/implémentation. Elle est prise en charge au sein de « lots » impliquant chercheurs en didactique, enseignants, concepteurs, développeurs. Les lots sont les suivants :

[1] Voir par exemple (Luste-Chaa, 2009 : 2.4) pour une présentation historique de la place du lexique en FLE.

[2] http://www.education.gouv.fr/cid206/les-langues-vivantes-etrangeres.html#Au_lycée___général,_technologique_et_professionnel

- SELF (Système d'Évaluation en Langues à visée Formative) vise à mettre en œuvre une méthodologie de création de tests de positionnement et à fournir la plate-forme permettant de les accueillir pour six langues (anglais, italien, mandarin, japonais, espagnol, français langue étrangère).
- Innovason regroupe les activités issues de :
    - THEMPPO (Thématique Prosodie & Production Orale) qui visait à proposer des outils et des pratiques centrées sur le développement des compétences prosodiques de la production orale en langue étrangère.
    - et COCA (Compétence Orale : Conception et Assistance) qui se fixait pour objectif de développer des solutions technologiques pour assister l'enseignant dans la conception d'activités de compréhension de l'oral.
- PARCOURS est centré sur le développement d'un module de l'ENPA (plate-forme fondée sur Claroline Connect) destiné à accueillir les parcours et réalisations des autres lots, mis à disposition de l'ensemble des acteurs de la formation.
- GAMER (Gaming Application for Multilingual Educational Resources) a pour objectif premier de concevoir et développer des jeux pour l'enseignement/apprentissage des langues, mais aussi de proposer des formations de formateurs sur l'usage du jeu en classe de langues.

Dans une seconde phase, les « équipes langues » (anglais, espagnol, italien, japonais, mandarin) s'approprient les outils conceptuels et technologiques conçus par les lots pour développer des contenus de formation. Pour organiser les développements autant que pour laisser mûrir les réflexions, les changements de phases ne sont pas complètement synchronisés (la seconde phase du lot SELF ne commence pas en même temps que celle du lot GAMER) et les phases ne sont pas hermétiques : équipes « lots » et « langues » collaborent durant les deux phases. Ces collaborations peuvent faire émerger un besoin différent et un chantier est alors créé pour explorer ce besoin. LexInnova s'inscrit dans ce type de travaux : il s'agit d'un chantier émergent visant à proposer des solutions concernant des enjeux partagés par plusieurs lots. L'équipe responsable de ce chantier est constituée d'enseignants de langue, de didacticiens et d'informaticiens spécialistes du TAL, tous issus de différentes équipes, d'Innovalangues et d'ailleurs.

Au sein du projet Innovalangues, les attentes en termes d'outillage du lexique s'expriment différemment : pour certains, les besoins principaux portent sur le contenu informationnel (voir par exemple Innovason, pour qui la transcription phonétique des mots du lexique semble primordiale), alors que d'autres utilisateurs potentiels sont plus tournés vers les fonctionnalités offerte par le système. Par exemple, pour le chantier Kinéphones, l'une des demandes concernait « l'écriture en couleurs » : une représentation différente des transcriptions phonétiques alignées sur la graphie où chaque phonème est associé à une couleur sur le modèle affranchi des contraintes éditoriales matérielles des tableaux de mots de l'approche Silent Way de Caleb Gattegno dématérialisé par Kinéphones[3].

3 Recueil des besoins

Après avoir recueilli les informations linguistiques pertinentes pour les collègues du projet par le biais d'une page du wiki interne d'Innovalangues, nous avons décidé, dans une perspective de « techniques » de conception centrées utilisateur (Bastien et Scapin, 2004 : 461), d'élaborer conjointement des scénarios d'utilisation visant à préciser les besoins fonctionnels. Ceux-ci s'appuient sur des cas pratiques et peuvent intégrer des éléments biographiques exemplifiant les interactions entre langue « maternelle » et langue « cible ».

[3] http://kinephones.u-grenoble3.fr/#/en_gb/1/table

3.1 Espagnol

Dans l'exemple de scénario suivant, légèrement retravaillé pour que le lecteur puisse en comprendre les enjeux, nous tentons de montrer comment la structure des lexiques et les besoins fonctionnels peuvent être intégrés à un scénario.

Anne étudie l'espagnol débutant depuis deux mois en LANSAD. Son enseignant a choisi de faire créer un lexique de groupe à tour de rôle par des étudiants en tandem. Pour ce faire, cette semaine, Anne et son partenaire travaillent sur l'ENPA. Ils relisent les notes prises durant le cours à la lumière des ressources proposées par l'enseignant et choisissent les mots qu'ils veulent inclure dans le lexique de groupe (d'un simple clic, quand la ressource est intégrée à l'ENPA). Dans certains cas, ils ajoutent à la notice de l'entrée un commentaire qui a été fait pendant la classe (ex : « Attention, en espagnol, '*padres*' signifie non seulement '*pères*', mais aussi '*parents*', ex : "*Llamo a mis padres cada semana*." ; (voir aussi hermanos, tíos) » dans l'article '*padre*').

Ils profitent également de ce travail à destination du groupe pour mettre à jour leur propre lexique et cliquent sur les mots qu'ils souhaitent rajouter dans leur lexique respectif (lexique personnel qui alimente le lexique de groupe et *vice-versa*). Anne souhaite par exemple s'approprier le mot « tío » (oncle). Elle clique sur ce mot, un menu déroulant s'ouvre. Elle peut alors consulter la ou les définitions ainsi que les exemples correspondant à ce mot (structure), puis importer ces données telles quelles ou en les modifiant. Anne décide par exemple de modifier l'exemple associé à « tío » pour son lexique personnel ; son oncle s'appelant Bernard, pour retenir « tío » elle veut l'associer à « Bernard ». Elle va donc choisir de mettre comme exemple personnel « Mi tío se llama Bernard » (besoin fonctionnel). Anne va également intégrer dans son lexique personnel des mots pris en notes durant le cours et durant l'écoute des enregistrements audio mis à disposition dans le cadre de sa formation. Anne, pour réviser le lexique de groupe et son lexique personnel, peut lancer des générateurs d'exercices (lien avec l'ENPA). La semaine suivante, un autre tandem se chargera du lexique de la semaine.

Eunice, bilingue français-portugais, est inscrite dans le même cours. Elle va consulter les mots nouvellement arrivés dans le lexique de groupe et faire des exercices à partir de ces mots (besoin fonctionnel). Elle choisira de rajouter une information spécifique dans son lexique personnel « Attention : accent sur le "i" de "tío" » (contrairement au portugais).

3.2 Mandarin

Les scénarios pour un lexique aussi distant que celui du mandarin ne diffèrent fondamentalement pas de ceux conçus pour les autres langues. En revanche, il existe un besoin impérieux que la prise en charge du lexique puisse au-delà — ou plus précisément en-deçà — de l'entrée dans le lexique, traiter un certain nombre d'éléments tels que les sinogrammes dont chaque entrée est constituée. En effet, en sus de la compétence lexicale au cœur de notre démarche, la compétence graphique est également traitée à travers des informations relatives à la combinatoire des sinogrammes, et à un niveau encore inférieur incluant les composants graphiques, les traits ou encore les différentes lectures sino-xéniques des sinogrammes dont sont composées les entrées du lexique.

Ainsi, Pierre, apprenant en mandarin, pourra accomplir les mêmes opérations que celles décrites plus haut.家 *jiā* – « famille », « maison » – vu en classe, figurera parmi les entrées de son lexique personnel. Plus tard, au détour d'une ressource en compréhension écrite, 國 家 *guojiā*, « pays » portera l'information complémentaire selon laquelle cette entrée est programmée au cours de la formation, le système étant capable d'intégrer le lexique des unités à venir – le *lexique cible* – incitant ainsi les apprenants à s'y intéresser au plus tôt.

L'apprenant pourra également confronter les sinogrammes de 國 家 *guójiā* à d'autres lexiques compilés par des institutions, qui associent à un niveau donné une liste de sinogrammes choisis selon différents principes. En effet, le paradigme didactique dominant du mandarin langue étrangère établit le sinogramme comme l'unité didactique fondamentale (Bellassen, 2010) induisant une chronopédagogie qui programme l'enseignement-apprentissage des sinogrammes selon un double

critère de fréquence d'un sinogramme dans le corpus et de la haute combinabilité de ce même sinogramme dans le lexique contemporain. Le corollaire d'une telle pratique fige des listes de sinogrammes par niveau qui constituent le socle de l'industrie certificative. Si l'approche didactique retenue pour Innovalangues veut rompre avec ce paradigme, nous ne pouvons pas exclure ces listes que nous étiquetons dans le système comme le *lexique institutionnel* informant l'utilisateur des coordonnées de chaque sinogramme dans différents contextes institutionnels : programmes de langue vivante de l'Éducation nationale, niveaux du test certificatif de chinois HSK etc.

Plus spécifique aux langues sinogrammiques en se distinguant du paradigme évoqué ci-dessus, LexInnova viserait à offrir une vue sur le lexique qui indiquerait les entrées qui partagent un même sinogramme. Ainsi, avec notre exemple, 家 apparaît également dans les entrées 作 家 *zuòjiā* « écrivain », 學 家 *xuéjiā* « universitaire », 文 學 家 *wénxuéjiā* « spécialiste de littérature » etc. ou encore 家 庭 *jiātíng* « famille », 家 族 *jiāzú* « clan » etc. Cette vue sur le lexique permet ainsi de traiter les dimensions suivantes : la polysémie de 家, qui en plus de vouloir dire « famille » renvoie par ailleurs à la notion d'agent lorsque qu'il se situe en dernière position. Il est même indiqué que la position de 家 dans sa première combinatoire 國 家 *guójiā* « pays » est une exception : en règle générale, et largement plus représentée dans le corpus, 家 renvoie à l'idée d'agent même si le paradigme communicationnel le présentera avant tout comme signifiant « famille ».

Au-delà de la mise en relation des différentes entrées contenant un même sinogramme, idéalement le système permettrait d'accéder aux informations suivantes : 家 est un sinogramme dont la structure est binaire verticale ⿱, constitué des composants 宀 *miǎn* « le toit », et 豕 *shǐ* « le porc ». Il est même possible d'intégrer une notice grammatologique infirmant l'édifiante étymologie orientaliste selon laquelle, pour les « Chinois des temps anciens, la famille était représentée au moyen d'un cochon sous le toit » alors que depuis le IIe siècle EC, nous savons que 家 s'est d'abord noté 豭 au moyen du composant du 豕 « porc » et d'un autre composant phonétique – 叚 *jiǎ* – encore quasi-homophone vingt siècles plus tard. A la faveur d'un processus orthographique inévitable sur une période historique aussi longue, le composant phonétique est tombé au profit d'un *composant discriminant* connu également sous le terme de « clé » : celui du toit 宀.

Enfin, le sinogramme est recontextualisé avec les sinogrammes qui partagent le même composant phonétique ou *phonophore* tel que 嫁 *jià* « prendre pour mari » ; mais également ses homophones parfaits ou au ton près tel que 加 *jiā* « ajouter » ou 價 *jià* « valeur, prix », 假 *jià* «vacances » ou *jiǎ* « faux »– qui partage le même phonophore historique de 家 *jiā* – et les liens vers les entrées contenant ces sinogrammes. Finalement, 家 *jiā* est recontextualisé en intercompréhension avec les lectures sino-xéniques et d'autres langues sinitiques, *ka* – ou plus rarement *ke* – en sino-japonais, *ga* en sino-coréen, *gia* en sino-vietnamien, et même en français à travers l'ethnonyme *hakka* – 客 家 *kèjiā* – désignant la communauté culturelle des vallées de l'arrière-pays des provinces méridionales du Fujian et du Guangdong, très présente dans la diaspora et majoritaire dans les confettis insulaires de l'empire français[4].

Toutes ces fonctionnalités et vues sur le lexique permettent ainsi de sous-traiter un maximum d'informations à la plate-forme et ainsi consacrer le plus possible du cours en présentiel à l'interaction orale plutôt qu'aux explications de vocabulaire et autres digressions graphiques et culturelles qui rendent le mandarin une langue encore plus distante pour les apprenants. Par ailleurs, une telle prise en charge du lexique permet de modéliser le profil de l'apprenant pour la génération d'exercices ainsi que les jeux évoqués ci-dessus.

[4] Le même changement phonétique d'une occlusive vélaire à une palatale est observable en français à travers l'emprunt Pékin au XVIIe siècle devenu depuis Beijing...

3.3 Relations avec GAMER

Les deux exemples précédents montrent les liens entre l'ENPA et les lexiques dans deux langues. Le cas du mandarin illustre plus spécifiquement les enjeux liés à la forme écrite (les sinogrammes) des mots-forme, alors que le cas de l'espagnol, il s'agit de l'exploitation à partir des entrées lexicales (lemmes ou mots-vedettes). C'est ce type de lien avec l'ENPA qui est au cœur de la problématique du lexique dans le lot GAMER. Dans le cadre de ces travaux, plusieurs prototypes de jeux sont en cours de développement avec un point de vue différent sur le lexique. Par exemple, dans le cadre du jeu « Magic Word[5] », dont les règles le placent dans la famille de jeux de lettres que l'on pourrait exemplifier par le Boggle (Loiseau, Zampa et Rebourgeon, 2015), il est nécessaire d'avoir une ressource lexicale qui définisse le champ des possibles (liste des formes autorisées). Ce jeu est focalisé sur ce que Portine appelle le pôle *accuracy*, c'est-à-dire sur le respect des schémas lexicaux et grammaticaux (Portine, 2013 : 162). Chaque langue prise en charge devra fournir une telle ressource contenant non-seulement « toutes » les formes de la langue, mais également les traits morphologiques associés et le lemme dont la forme est issue : certaines règles des futures versions du jeu en dépendent. Mais un autre besoin est la possibilité de lier les lemmes (associés aux formes trouvées dans le jeu) à l'activité du joueur dans les parties plus formelles de l'ENPA. Dans ce cas-là les générateurs d'exercices mentionnés dans le scénario espagnol pourraient être lancés à partir de cette liste de mots (dans ce contexte, le lexique du joueur, constitué automatiquement d'après les traces d'interaction ferait office de profil de l'apprenant).

Un autre exemple de jeu, radicalement différent dans ses objectifs, est « Game of Words[6] » (Loiseau, Hallal et Ballot, 2016). Ce jeu de « devinettes » est lui tourné vers le pôle *fluency*, centré sur les activités discursives (Portine, 2013 : 162). Entre autres activités attendues du joueur, il s'agira ici d'effectuer un enregistrement audio permettant de faire deviner un mot aux autres joueurs en respectant des règles énonciatives (dans la première version, celles du jeu *Taboo*), ce qui demande d'avoir une ressource lexicale qui associe un mot à d'autres entrées utiles pour le définir (mots interdits). Dans le cadre de ce jeu, l'une des passerelles anticipées est la possibilité pour un joueur de privilégier les mots qui figurent déjà dans son dictionnaire personnel afin de lui fournir un autre point de vue sur celui-ci. Par exemple, « Anne » pourrait demander à jouer sur les mots de son lexique personnel.

Enfin, un dernier exemple de jeu et d'intégration spécifique serait celui de *Kanji Crunch*, un jeu d'association, dont il n'existe pas de prototype à l'heure actuelle contrairement aux deux autres. Celui-ci vise à permettre au joueur d'appréhender globalement l'économie du système graphique en mandarin (Goudin et Lê, 2016). Une fois les premiers niveaux joués et les compétences de base acquises (Chaix et al., 2016), un système comme LexInnova permettrait au joueur de se confronter à des niveaux construits à partir des sinogrammes issus de son lexique de groupe (cf. scénario espagnol et section 4.2.3). Mais, par le biais de cette question du sinogramme, le mandarin pose également des questions spécifiques quant à la structure du lexique.

Conclusion

À travers les exemples proposés ci-dessus, nous avons tenté de montrer d'une part la diversité des besoins en termes de structuration du lexique, mais également les ponts qui pourraient être dressés entre différents travaux menés indépendamment du fait des fonctionnalités nécessaires à la prise en compte de ces besoins. Ceux qui ont été présentés ici sont avant tout monolingues, mais d'autres (voir par exemple le projet Check Your SMILE (Yassine-Diab, Alazard-Guiu et Loiseau, 2016)), présentent également des besoins de traduction.

Au-delà des exemples ci-dessus, la plupart des équipes du projet Innovalangues développe des contenus dans plusieurs langues, ce qui nécessitent tout un outillage pour la gestion du lexique. Dans la suite de l'article, nous présenterons nos propositions pour un tel outil.

[5] http://gamer.innovalangues.net/magicword
[6] http://gamer.innovalangues.net/gameofwords

4 Conception d'une base lexicale multilingue

Bien que la gestion des développements et déploiement de l'ENPA, et donc la coordination des développements informatiques, soit centralisée, les produits issus de chaque lot sont conçus et développés indépendamment, le lien se faisant par l'intégration à l'ENPA. Mais les ressources linguistiques comme les lexiques et les corpus, d'une très fine granularité, sont structurés de façon complexe et très spécifique, et leurs tailles posent le problème du passage à l'échelle, qui doit être traité au niveau de chaque chantier.

Le projet Innovalangues est multilingue, puisqu'il permet aux élèves de s'inscrire et gérer leur formation dans plusieurs langues. Chaque langue peut sembler apprise indépendamment quand on ne considère que l'inscription d'un élève à un parcours. Mais le profil de l'élève et son histoire personnelle révèlent les spécificités (caractéristiques) multilingues de son rapport aux ressources linguistiques. De plus, les travaux en intercompréhension menés de longue date au LIDILEM (Carlo et al., 2015 ; Dabène, 1994) et intégrés aux formations du LANSAD[7] soulignent à nouveau les besoins de ressources lexicales multilingues centralisées.

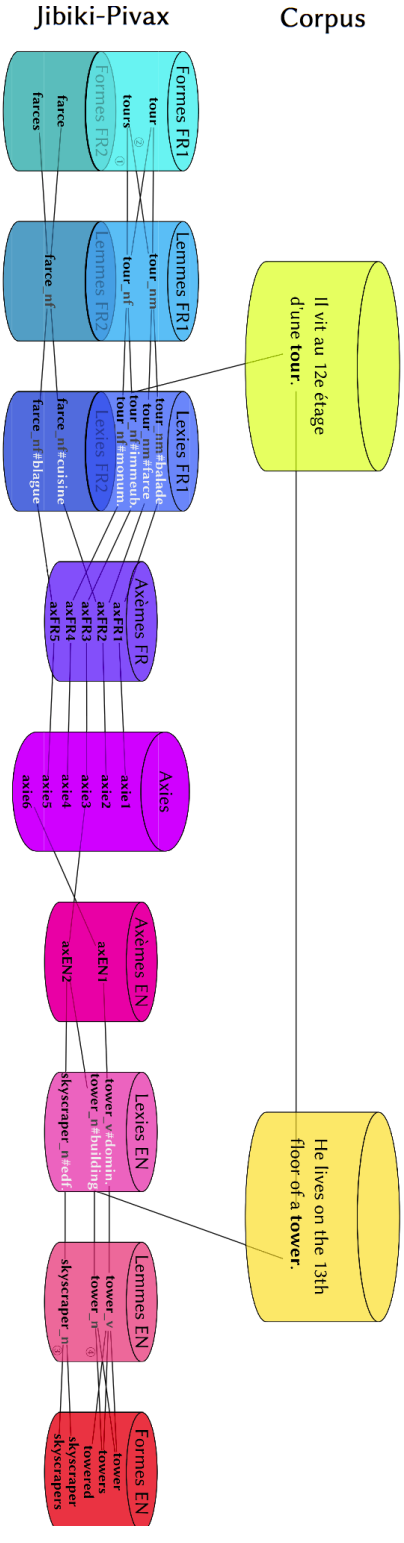
4.1 Macrostructure étendue

Les lots Innovason et GAMER exploitent tous un ou des lexique(s) réalisé(s) dans une base de données dédiée interne au lot. La diversité des besoins aurait demandé un long travail de coordination qui aurait trop ralenti chacun des chantiers, mais à l'heure où la question d'un outil centralisé pour la gestion du lexique se pose, la collecte des microstructures de chaque lexique (formes fléchies, lemmes et traits morphologiques associés[8], transcription phonétique[9], définitions L1 et L2[10] , [11], syllabes, accentuation[12], sinogrammes[13], exemples en contexte[14], etc.) a conduit à concevoir une structure de base lexicale multivolumes (cf. figure 5).

Les besoins d'accès par la forme fléchie conduisent à l'identification d'un volume pour celles-ci. L'intégration des transcriptions phonétiques doit être reliée aux formes fléchies. Cela conduit aussi à l'identification d'un volume dédié.

[7] http ://lansad.u-grenoble3.fr/version-francaise/formations-en-langues/intercomprehension-en-langues-romanes-77120.kjsp

[8] cf. Magic Word
[9] cf. InnovaSon
[10] cf. le scénario espagnol ci-dessus
[11] cf. GAMER
[12] cf. InnovaSon
[13] cf. scénario mandarin ci-dessus
[14] cf. le scénario espagnol ci-dessus



Prenons l'exemple de la graphie "tour" en français. Sa transcription phonétique sera /tur/. La forme considérée ("tour") peut-être un nom masculin, "tour_nm", ou un nom féminin, "tour_nf", gérés dans le volume dédié aux lemmes. Mais ces deux lemmes peuvent avoir des sens différents, que l'on peut préciser avec une glose, comme :

1. tour_nm#potier, tour_nm#balade, tour_nm#perimètre, tour_nm#magic, tour_nm#farce,

2. tour_nf#contrôle, tour_nf#PC, tour_nf#immeuble, tour_nf#monument.

Ces différents sens sont portés par les lexies, qui sont gérées par un volume dédié. Une lexie est donc liée à une forme pour porter un sens précis.

Il est possible qu'un autre lemme conduise au même sens. Sa lexie sera différente mais le sens "unique". Les sens partagés par plusieurs lexies sont les axèmes. Un volume dédié les stocke. Disons que dans notre exemple, "tour_nm#immeuble" soit identifié par l'axème "id-axm-3856900".

Une fois le sens monolingue ainsi précisé, pour passer au même sens dans les autres langues (et pas seulement une autre langue), on définit les axies. Formellement, une axie est une classe d'équivalence de lexies synonymes (Boitet, Mangeot, Sérasset, 2002).

Si notre axème "id-axm-3856900" est relié à l'axie "id-axi-98663940", la liaison de cette axie avec les axèmes de l'anglais, gérés dans le volume des axèmes de l'anglais, pourrait conduire à identifier l'axème "id-axm-1001123", qui est lié à la lexie "tower_n#building" mais aussi à la lexie "tower_n#skyscraper". Ces lexies conduisent à deux gloses : les noms "tower_n" et "skyscraper_n". Chacun de ces noms ont deux formes, l'une au singulier, l'autre au pluriel. Les différents volumes sollicités sont présentés dans la figure 5.

Figure 1: Proto macrostructure de la base lexicale

On peut créer une telle base lexicale dans un logiciel comme Jibiki, mais :

- d'une part, il faudra étendre les fonctionnalités offertes par l' API ;
- et d'autre part, les requêtes reliant une forme dans une langue avec les formes correspondantes dans une autre langue seront coûteuses en volume. En effet, d'un point de vue conceptuel au sens des modèles entités-associations pratiqués en conception de bases de données, les lexies réalisent une association n-m entre les lemmes et les axèmes, de même pour les axies entre les axèmes des langues considérées. Il y a encore des associations n-m entre les lemmes et les formes, et entre les formes et les transcriptions phonétiques.

Les associations n-m augmentent les possibilités de parcours de façon multiplicative, même si certaines peuvent se compenser. Il en résulte un risque de temps de réponse intolérable pour les utilisateurs dans le cadre de leur tâche d'apprentissage, pour des requêtes sollicitant la base lexicale dans toute la largeur de ses volumes. Toutefois, les régularités de la langue diminueront probablement grandement la combinatoire et on pourra travailler sur des sous-volumes pour minimiser les tailles intermédiaires.

4.2 Portée des différents types de ressources

Les scenarii proposés mettent en relief des statuts différents de certaines données de la base lexicale :

- données relatives à l'utilisateur (choisies explicitement ou collectées du fait de son activité) ;
- données relatives à des groupes d'utilisateurs ;
- données issues de ressources externes.

Ceci nous a forcés à formaliser ces statuts, définissant ainsi une macrostructure fonctionnelle pour notre prototype.

4.2.1 Dictionnaire de référence

Le choix du terme **dictionnaire de référence** provient des entités (objets) qui serviront à peupler le contenu de la base lexicale. Par transitivité, on pourra parler de « dictionnaire de référence » (concept) pour parler des informations contenues dans la base lexicale et issues des dictionnaires de référence-objets. Dès lors, se pose la question du statut des informations présentes dans la base lexicale qui ne sont pas issues des dictionnaires de référence (objets). Il semble cohérent de considérer les modifications de la base lexicales validées comme intégrées au dictionnaire de référence (concept). On en arrive à la définition suivante :

dans le cadre de LexInnova, le dictionnaire de référence est la somme des informations de la base lexicale, qui a été validée, que ce soit par le biais d'une institution externe à l'instance de LexInnova (cf. dictionnaires existants) ou par le biais du système.

4.2.2 Lexique institutionnel

U n **lexique institutionnel** a pour vocation de synthétiser le point de vue sur le lexique d'une institution. Une institution pourra être tout à fait externe à LexInnova (ex : Conseil de l'Europe) ou un groupement d'utilisateurs de LexInnova. Ce lexique constitue une forme de référentiel, il n'a pas vocation à évoluer selon le déroulement d'un cours. On pourra en revanche créer un lexique institutionnel qui traduise le déroulement générique de cours dans une institution (cf. lexique cible dans le scénario « mandarin »).

4.2.3 Lexique de groupe

U n **lexique de groupe** est un lexique dynamique qui peut être modifié par un ensemble d'utilisateurs. La configuration de base sera un enseignant propriétaire d'un lexique de groupe, qui y donne accès (y compris en écriture) à un groupe d'apprenants. Toutefois, on peut tout à fait imaginer un groupe d'apprenants se créant entre eux un lexique de groupe (par exemple pour l'argot). Cela posera à terme la question de la visibilité du lexique : un lexique de groupe doit-il être visible pour les autres groupes ? Un lexique créé par un groupe d'apprenant doit-il être consultable par les enseignants ?

Le lexique de groupe permet également de formuler des objectifs pour les membres du groupe. Chaque entrée d'un lexique de groupe pourra être taguée avec la valeur cible. Ainsi le lexique cible d'un groupe pourra évoluer tout au long de l'année en fonction des ajouts de ses membres ou d'imports issus de lexiques institutionnels ou personnels.

4.2.4 Lexique personnel

U n **lexique personnel** est un lexique qui n'a qu'un seul propriétaire. Il s'agira en général d'un apprenant, qui seul pourra modifier son contenu et décider de sa visibilité. Ici aussi le propriétaire du lexique pourra choisir les éléments cibles du lexique. Au sein de ce lexique personnel, les données pourront être étiquetées pour faire la différence entre les entrées que l'apprenant a explicitement ajoutées à son lexique et celles qui y ont été ajoutées pour une raison ou pour une autre par le système (cf. scénarios « jeu »). Quoi qu'il en soit, l'apprenant membre d'un groupe devra décider lui-même si ses objectifs coïncident avec ceux du groupe (des groupes) au(x)quel(s) il appartient. En d'autres termes, ce n'est pas parce qu'une entrée figure dans un lexique de groupe auquel a accès un apprenant qu'elle se trouvera dans son lexique personnel.

5 Résultat fonctionnel

Si les étapes précédentes nous ont permis d'affiner la modélisation du lexique, des questions restent en suspens (cf. partie précédente). Afin d'alimenter les discussions avec les futurs utilisateurs et d'orienter nos recherches, nous avons décidé d'implémenter rapidement un prototype fonctionnel. Celui-ci sert également à décider des efforts de développement futurs. Ce prototype est développé en PHP de façon à pouvoir être intégré facilement à l'ENPA développé dans le projet Innovalangues. Il se sert de la plate-forme Jibiki pour la gestion des ressources lexicales. La communication entre les deux systèmes s'effectue via l'interface de programmation de Jibiki fondée sur le protocole REST.

Jibiki (Mangeot & Chalvin, 2006) est une plate-forme générique de gestion de ressources lexicales hétérogènes en ligne développée et maintenue depuis 2001 principalement par Mathieu Mangeot. Celle-ci est programmée à l'aide de Enhydra, un serveur d'objets java basé sur une architecture 3/tiers. Pour la couche de données, la base de données utilisée est Postgres. Le logiciel est disponible gratuitement et en source ouverte sur la forge du LIG[15].

Tout type de ressource lexicale au format XML peut être importé dans la plate-forme. Les ressources sont décrites et manipulées à l'aide de pointeurs dans la structure XML, ce qui permet de les importer, les consulter et les éditer sans modifier leur structure. Il est possible de gérer tout type de microstructure. Concernant la microstructure, le système de traitement de liens inter-volumes permet de gérer différentes macrostructures : monolingue, bilingue bidirectionnelle, multilingue à structure pivot ou multi-étages type Pivax (Zhang et al., 2014).

5.1.1 Consultation des ressources

Deux interfaces de consultation sont disponibles. L'interface de consultation simple affiche une vue évoluée du dictionnaire imprimé : la partie gauche affiche les vedettes du voisinage immédiat du

[15] https : //jibiki.ligforge.imag.fr/

mot recherché classées par ordre alphabétique. Elle est équipée d'un ascenseur infini permettant de parcourir toutes les vedettes du dictionnaire selon l'ordre alphabétique. Lorsqu'on clique sur une vedette de la partie gauche, l'article complet s'affiche dans la partie droite.

L'interface de consultation avancée permet de combiner les critères de recherche. Par exemple, rechercher un mot français terminant par "er", qui soit un nom et qui appartienne au domaine de la botanique.

Lors de l'affichage de chaque article, un menu s'affiche en haut à droite. Celui-ci comprend des liens vers le formulaire d'édition, l'historique des modifications, ainsi que la vue source XML de l'article.

5.1.2 Édition des articles

Lors de l'import d'une ressource sur la plate-forme, le schéma XML représentant la structure des articles est également importé. Un formulaire d'édition des articles est ensuite automatiquement généré par la plate-forme (Mangeot & Chalvin, 2006).

Deux modes d'édition sont possibles. Lors de la consultation d'un article, l'édition sur place permet de modifier directement le texte d'un article qui s'affiche en double-cliquant sur la chaîne de caractères à modifier. Celle-ci se transforme en champ de texte avec un bouton « ok »sur la droite pour valider la saisie.

Cet éditeur est programmé à l'aide de la technologie AJAX et il utilise l'API REST de la plate-forme Jibiki pour dialoguer avec le serveur. Lors de la validation de la saisie (clic sur le bouton « ok »), la nouvelle chaîne de caractères est envoyée au serveur avec le pointeur Xpath du segment édité et l'identifiant unique de l'article (Mangeot, 2016).

Par contre, ce mode ne permet pas l'ajout de nouvelles informations, comme un exemple d'usage en contexte. Pour ajouter ou supprimer des parties d'information, il est nécessaire d'utiliser le formulaire d'édition complète. Outre les interacteurs HTML classiques (boîtes texte, menu déroulant, cases à cocher, etc.), celui-ci a été enrichi avec des interacteurs plus complexes pour gérer des listes d'objets.

5.2 Présentation du système Lexinnova

5.2.1 Architecture globale du système

Le site Web principal avec lequel l'utilisateur interagit est l'ENPA du projet Innovalangues.

L'ENPA est déjà intégré à la plate-forme Claroline Connect. Le prototype Lexinnova [16] reprend la charte graphique de l'ENPA et il est en cours d'intégration sous forme de module Claroline Connect.

Les fonctionnalités du prototype sont directement intégrées à l'ENPA sous forme de menus en haut à droite de l'écran.

Une instance spécifique de Jibiki a été déployée pour le projet Lexinnova.

Celle-ci contient :

- un dictionnaire de référence Lexinnova avec un volume pour chaque langue (pour l'instant seul l'espagnol est accessible) ;
- un dictionnaire créé pour chaque utilisateur avec un volume pour chaque langue.

Cette instance est installée sur le serveur de démonstration de l'équipe GETALP, au Laboratoire d'Informatique de Grenoble. La communication entre le prototype et l'instance de Jibiki s'effectue avec l'interface de programmation (API) REST de la plate-forme Jibiki. La figure 2 montre l'architecture du système dans son ensemble.

[16] http : //totoro.imag.fr/Lexinnova

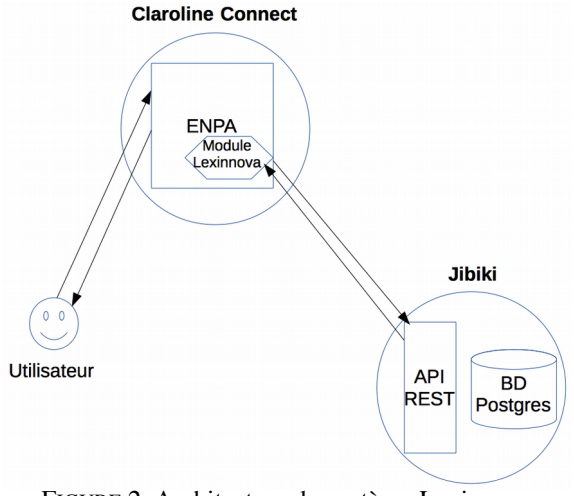


FIGURE 2: Architecture du système Lexinnova

5.3 Fonctionnement du prototype

Lors de la première visite d'un utilisateur sur l'ENPA et avant que son compte ne soit créé, deux fonctionnalités sont disponibles dans le menu Lexinnova :

- consultation des dictionnaires de référence ;
- lecture active.

Lorsque l'utilisateur crée un compte sur l'ENPA, un compte similaire est créé sur la plate-forme jibiki via un appel à l'API.

Une fois le compte créé, une nouvelle fonctionnalité est ajoutée au menu : la gestion des lexiques personnels (consultation et modification).

Si l'utilisateur a des droits spécifiques d'administration (par exemple, un enseignant ou un administrateur de l'ENPA), un second menu d'administration est ajouté à la barre des menus.

Il contient les fonctionnalités suivantes :

- gestion des utilisateurs (liste de tous les utilisateurs et suppression d'un utilisateur existant);
- gestion des lexiques (liste de tous les lexiques, création d'un nouveau lexique et suppression d'un lexique existant).

Une démonstration de ce prototype a été réalisée lors du séminaire bi-annuel Innovalangues de janvier 2016[17]. Un podcast est à présent disponible[18].

5.4 Fonctionnalités disponibles

5.4.1 Limitations du prototype

De façon à pouvoir rapidement montrer un prototype fonctionnel, nous avons choisi de limiter les fonctionnalités disponibles en nous concentrant sur certains scénarios (scénario « tío »). La réflexion

[17] http://innovalangues.fr/seminaire-bi-annuel-innovalangues-janvier-2016/
[18] http://podcast.grenet.fr/episode/lexinnova-lexiques-personnalisables-integres-a-lenpa/

sur les lexiques de groupe étant encore en cours, nous n'avons pas implémenté cette fonctionnalité. Les dictionnaires de références et les lexiques personnels ont été quant à eux complètement implémentés.

Au niveau des données, nous n'avons pas encore collecté un ensemble complet de données lexicales pour chaque langue du projet. Nous avons créé un dictionnaire de référence pour l'espagnol afin d'implémenter certains scénarios. Ce dictionnaire ne contient pour l'instant que le vocabulaire nécessaire à la réalisation des scénarios (soit une vingtaine de mots au total).

Pour le passage à l'échelle en termes de quantité de données et de nombre d'utilisateurs, des calculs ont été faits sur la base de 180 000 entrées pour chaque dictionnaire de référence et 10 000 utilisateurs inscrits. Des tests de montée en charge ont été effectués sur la base de 600 utilisateurs en ligne. Le prototype actuellement déployé peut répondre à une telle charge pour les fonctionnalités principales de gestion des ressources. Par contre, le module de lecture active nécessitant beaucoup de mémoire vive, il sera nécessaire de le déployer sur un autre serveur afin de répartir la charge.

5.4.2 Consultation des dictionnaires de référence

La consultation des dictionnaires de référence est publique. Il n'est donc pas nécessaire de se loguer pour y accéder. L'interface utilisée est la page de consultation simple de la plate-forme Jibiki : la partie gauche affiche les vedettes du voisinage immédiat du mot recherché et la partie droite affiche le ou les articles trouvés de manière détaillée. Si l'utilisateur est logué, un bouton "+" s'affiche en haut à droite de l'article. L'utilisateur peut cliquer sur ce bouton pour importer l'article dans son lexique personnel.

La figure 3 montre la fenêtre de consultation du mot "tio" dans le dictionnaire espagnol (on notera

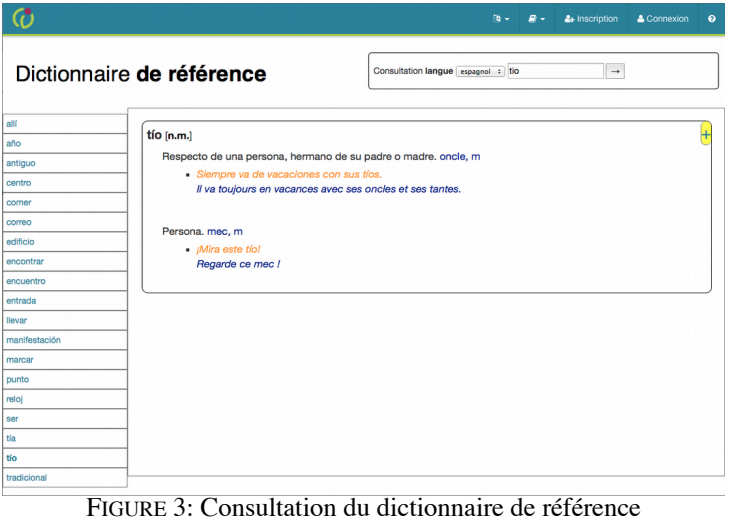
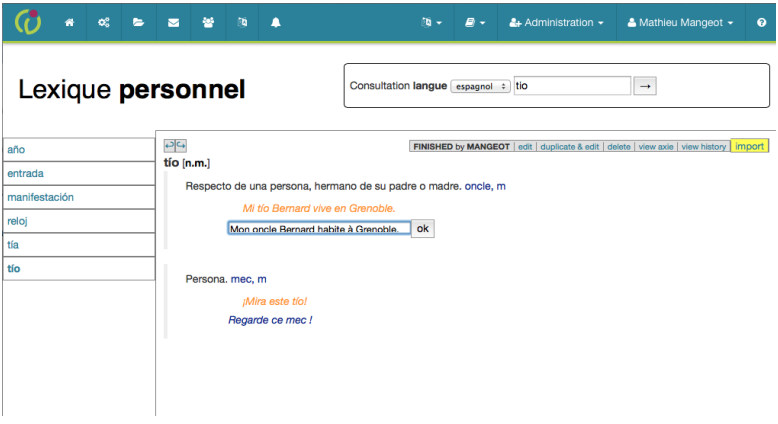


FIGURE 3: Consultation du dictionnaire de référence

que l'article « tío », accentué, est obtenu).

5.4.3 Gestion des lexiques personnels

FIGURE 4: Édition d'un exemple d'article dans le lexique personnel5



La gestion des lexiques personnels reprend également l'interface de consultation simple de la plate-forme Jibiki. Il est par contre nécessaire d'être un utilisateur enregistré pour pouvoir y accéder. Lorsque l'utilisateur consulte une entrée dans cette interface, il peut ensuite l'éditer directement en cliquant sur la zone à modifier. La technique utilisée est celle de l'édition sur place de la plate-forme Jibiki. Sur la figure 4, l'utilisateur a consulté le mot "tío" dans son lexique personnel (après l'avoir importé depuis le dictionnaire de référence). Il souhaite modifier l'exemple pour qu'il fasse référence à son histoire personnelle ce qui lui permettra d'accéder plus facilement au sens de ce mot.

5.4.4 Module de lecture active

Le module de lecture active reprend celui utilisé sur le site jibiki.fr (Mangeot 2016). Il s'agit d'une aide à la lecture pour les apprenants d'une langue étrangère. L'utilisateur saisit ou colle un texte dans la fenêtre de saisie. Le texte est ensuite envoyé à un analyseur morphologique puis chaque lemme est consulté dans son lexique personnel ou dans le dictionnaire de référence si aucune réponse n'est trouvée dans son lexique personnel. Le texte entré précédemment est réaffiché plus bas avec des informations supplémentaires provenant de l'analyse. Lorsque l'utilisateur bute sur un mot, il clique dessus et une fenêtre s'affiche avec l'article correspondant provenant de la consultation des différentes ressources (lexique personnel ou dictionnaire de référence).

Dans l'exemple de la figure 5, l'utilisateur clique sur le mot "antiguo" et l'article s'affiche sur la droite de l'écran. Si l'article provient du dictionnaire de référence, celui-ci affichera un bouton "+" permettant d'importer l'article dans le lexique personnel de l'utilisateur, comme lors de la consultation du dictionnaire de référence.

Le module de lecture active peut également afficher une transcription ou une prononciation au dessus de chaque mot en utilisant l'élément ruby[19] du standard HTML5. Pour le japonais, il est possible d'afficher le furigana ou le romaji. Pour le chinois, il est possible d'afficher le pinyin ou le bopomofo. Pour le français, il est possible d'afficher une prononciation simplifiée ou en API, etc.

[19] http://www.w3.org/TR/ruby/

FIGURE 6: Module de lecture active

6 Conclusion

La conception d'un ENPA a fait émerger la nécessité d'une grande base lexicale commune couvrant tous les besoins des différents acteurs, que se soient des machines (modules de jeux sérieux ou de générateurs d'exercices, etc.) ou des humains (apprenants). À partir de plusieurs scénarios d'utilisation, nous avons défini une structure de base lexicale multilingue permettant de répondre à ces besoins. Nous avons ensuite implémenté un prototype fonctionnel permettant la consultation automatique ou manuelle de ressources et la création de lexiques personnels. Ce prototype, programmé en PHP, interagit avec la plate-forme Jibiki de gestion de ressources lexicales.

Les perspectives principales de ce projet à moyen terme sont les suivantes :

- l'analyse des pratiques d'enseignants en langues en termes de lexiques grâce à un questionnaire en cours d'élaboration ;
- le passage à l'échelle du prototype en termes de nombre d''utilisateurs, de quantité de données et de macrostructure. Pour cela, il faudra constituer pour chaque langue un volume de grande couverture avec des données libres de droit (possibilité d'utiliser les données de wiktionary) et de relier les différents sens de chaque langue via un volume pivot ;
- un test grandeur nature au deuxième semestre 2016-2017 avec une classe d'étudiants en sciences apprenant l'anglais de spécialité à l'Université de Grenoble.

Remerciements

Cette recherche a été en grande partie financée par le projet IDEFI Innovalangues.

Références

BASTIEN, C. ET SCAPIN, D. (2004). « La conception de logiciels interactifs centrée sur l'utilisateur: étapes et méthodes », *in* Pierre Falzon (dir.), *Ergonomie*, 1re édition, Paris, Presses Universitaires de France, pp. 451–462.

BELLASSEN, J. (2010). « La didactique du chinois et la malédiction de Babel : émergence, dynamique et structuration d'une discipline » in *Études Chinoises*, Hors-série. Disponible en ligne : http://www.afec-etudeschinoises.com/IMG/pdf/Bellassen_Didactique.pdf.

BOITET, C., MANGEOT, M., SÉRASSET, G. (2002), « The PAPILLON project: cooperatively building a multilingual lexical data-base to derive open source dictionaries & lexicon », Proc of on NLP and XML (NLPXML 2002), Ed. Graham Wilcock, Nancy Ide & Laurent Romary, COLING Workshop, Taipei, Taiwan, 31 August 2002, pp. 9-15.

CARLO, M.D., ANQUETIL, M., VECCHI, S., JAMET, M.-C., MARTIN, E., PEREA, E.C., HIDALGO, R., PISHVA, Y., GILLES, F. ET ANDRADE, A.I. (2015). « REFIC – Référentiel de compétences de communication plurilingue en intercompréhension », Référentiel de compétences Projet Européen Miriadi. Disponible en ligne : http : //www.miriadi.net/refic.

CARRON T., MARTY J-C. ET MANGEOT M. (2009). *How to bring immersion in Learning Games ?* Proc. IEEE-ICALT 2009, Riga, Latvia, 15-17 July 2009, 6 p.

CECRL (2000). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*, Conseil de la coopération culturelle — Comité de l'éducation — Division des langues vivantes (dir.), Édition française, Strasbourg; Paris, Conseil de l'Europe. Disponible en ligne : http : //medias.didierfle.com/media/contenuNumerique/007/4140016745.pdf.

CHAIX, C., MEUNIER-CARUS, M., PESQUET, J.-A., RONDIN, R., GOUDIN, Y. ET LOISEAU, M. (2016). « Kanji Crunch: Apprentissage dissocié des habiletés et autonomisation de la compétence graphique par le jeu vidéo en mandarin langues étrangères », 38e colloque APLIUT : *Jeux en jeu dans l'enseignement/apprentissage des langues en LANSAD*, Lyon, 2-4 juin 2016.

DABÈNE, L. (1994). « Le projet européen GALATEA: pour une didactique de l'intercompréhension en langues romanes » Jeanine Stolidi (dir.), *Recherches en linguistique hispanique*, n°22, pp. 41-45.

GOUDIN, Y. ET LÊ, T.M. (2016). « Jouer avec le sacré? Le sinogramme à l'ère du jeu sérieux » Haydée Silva et Mathieu Loiseau (dir.), *Recherches et applications*, n°59, pp. 145–160.

LOISEAU, M., HALLAL, R. ET BALLOT, P. (2016). « Game of Words: prototype of digital game focusing on oral production (and comprehension) through asynchronous interaction », EUROCALL 2016, Limassol (Chypre), 24-27 août 2016.

LOISEAU, M., ZAMPA, V. ET REBOURGEON, P. (2015). « Magic Word: premier jeu développé dans le cadre du projet Innovalangues » , *ALSIC*, vol. 18, n°2. DOI : 10.4000/alsic.2828. Disponible en ligne : http : //alsic.revues.org/2828.

LUSTE-CHAA, O. (2009). *Les acquisitions lexicales en français langue seconde: conceptions et applications*, Thèse de doctorat, Université Paul Verlaine, Metz. Disponible en ligne : http://www.theses.fr/2009METZ023L.

MANGEOT M. (2016). *Collaborative construction of a good quality broad coverage and copyright free Japanese-French dictionary* Hosei University International Found Foreign Scholar Fellowship Report Volume XVI 2013-2014, Hosei University, Tokyo, Japan, pp. 175-208.

MANGEOT M. ET CHALVIN A. (2006). *Dictionary Building with the Jibiki Platform : the GDEF case.* Proc. of LREC 2006, Genoa, Italy, 23-25 May 2006, pp. 1666-1669.

MASPERI, M. ET QUINTIN, J.-J. (2014a). « Enseigner à l'université en France, à l'ère du numérique: l'apport de dispositifs d'ingénierie innovants dans la formation en langues », *in* Cristiana Cervini et Anabel Valdivieso (dir.), *Dispositivi formativi e modalità ibride per l'apprendimento linguistico*, Bologna, CLUEB, Contesti Linguistici, pp. 61-80. Disponible en ligne : https ://www.researchgate.net/publication/271852910.

MASPERI, M. ET QUINTIN, J.-J. (2014b). « L'innovation selon Innovalangues » Elsa Del Col (dir.), *Lingua e nuova didattica*, n°1/2014, pp. 6-14.

PORTINE, H. (2013). « L'ingénierie linguistique : des technologies au service d'une didactique intégrant la cognition ? » Christian Ollivier et Laurent Puren (dir.), *Mutations technologiques, nouvelles pratiques sociales et didactique des langues*, n°54, pp. 159–168.

YASSINE-DIAB, N., ALAZARD-GUIU, C. ET LOISEAU, M. (2016). « Check your Smile, first prototype of a collaborative LSP website for technical vocabulary learning », EUROCALL 2016, Limassol (Chypre), 24-27 août 2016.

ZHANG Y., MANGEOT M., BELLYNCK V. ET BOITET C. (2014). *Jibiki-LINKS : a tool between traditional dictionaries and lexical networks for modelling lexical resources.* Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex) 2014 (Eds. Michael Zock, Reinhard Rapp, Chu-Ren Huang), Dublin, Ireland, 23 August 2014, 12 p.

Génération d'exercices d'apprentissage de langue de spécialité par l'exploration du corpus

François-C. REY, Izabella THOMAS, Iana ATANASSOVA
Centre L. Tesnière, Université de Franche-Comté
30 rue Mégevand 25030 Besançon, France
fcrey@edu.univ-fcomte.fr, izabella.thomas@univ-fcomte.fr,
iana.atanassova@univ-fcomte.fr

RÉSUMÉ

Il n'existe pas de travaux sur des logiciels dédiés à l'apprentissage des langues de spécialité, ni à la création de ressources pédagogiques spécifiques. Les outils informatisés concernant les langues de spécialité se limitent aux logiciels d'aide à la traduction et à des outils d'aide au recensement terminologique. Afin de doter les enseignants de langues étrangères de spécialité d'outils prenant en compte leurs besoins spécifiques nous proposons de concevoir et de développer une plateforme pour la génération de matériels pédagogiques dans le domaine des langues de spécialités. Pour répondre à cet objectif, nous avons effectué une première expérimentation concernant la génération automatique d'exercices d'apprentissage du vocabulaire spécialisé. Nous avons choisi la génération d'exercices à trous basés sur les phrases tirées d'un corpus de textes authentiques. Les résultats de ces premières expérimentations montrent la faisabilité de la génération d'exercices à trous, à partir de listes de vocabulaire et d'un corpus spécialisés intégrés à la plateforme, ainsi que de textes fournis par un enseignant.

ABSTRACT

A Corpus Based Approach to the Generation of Exercises for Language for Specific Purposes. There are neither works on software dedicated to the learning of language for specific purposes, nor works dedicated to the creation of specific educational resources. The computerized tools concerning the languages for specific purposes are limited to the computer aided translation software and to terminological census tools. In order to equip foreign language teachers with tools taking into account their specific needs, we propose to develop a platform for the generation of educational materials in the field of language for specific purposes. In this conceptual framework, we carried out a first study on the automatic generation of vocabulary lerning exercises for language for specific purposes. We chose the generation of cloze tests based on the sentences from a corpus of authentic texts. The results of these first experiments show the feasibility of the generation of cloze tests, starting from lists of vocabulary and from a corpus wich are specialized, integrated into the platform, and from texts supplied by a teacher.

MOTS-CLÉS : apprentissage des langues assisté par ordinateur - ALAO, enseignement des langues assisté par ordinateur - ELAO, langues de spécialité - LSp, génération d'exercices, exercices de vocabulaire, exercices en contexte, anglais de spécialité - ASP

KEYWORDS: Computer-Assisted Language Learning - CALL, contextual exercises, Language for Specific Purposes - LSP, English for Specific Purposes - ESP

1. Introduction

Les apprenants sont au cœur des systèmes logiciels d'apprentissage des langues assisté par ordinateur (ALAO) et d'enseignement des langues étrangères par ordinateur (ELAO). En revanche, les enseignants de langues disposent de peu d'outils répondant à leurs besoins en matière de préparation des activités d'enseignement. Pourtant, de nombreuses tâches peuvent faire objet d'aides automatisées, simplifiant ainsi le travail des enseignants, réduisant le temps de préparation de matériel, ou facilitant la création de ressources pédagogiques complexes à élaborer. Les besoins exprimés par les enseignants de langues vont au-delà de l'utilisation de logiciels génériques de mise en forme d'exercices. Un exemple d'un tel logiciel est *Hot Potatoes*[1], qui offre la possibilité de créer plusieurs types d'exercices sous forme de pages web interactives. Ce type de logiciels s'adapte à toutes les matières puisqu'il facilite uniquement la mise en forme et non pas la création des contenus des exercices. Cette dernière tâche est parmi les plus complexes lors de la préparation de matériels pédagogiques par des enseignants. L'objectif de notre travail consiste à intégrer des ressources et matériaux spécifiques utilisés en cours de langues.

La question qui nous intéresse particulièrement est celle de l'enseignement de langues de spécialité. Il pose des problématiques bien différentes de celle de l'enseignement de langues étrangères en général. Par langue de spécialité nous entendons un sous-ensemble d'une langue naturelle utilisé par un groupe de spécialistes à l'intérieur d'un domaine du savoir ou d'une discipline technique (définition inspirée de L'HOMME [L'HOMME, 1990]). Le domaine des langues de spécialité se confond au moins en partie avec celui des *langues pour (les) spécialistes d'autres disciplines* (LANSAD), et, dans une moindre mesure, avec celui du *Vocationally (and Professionally) Oriented Language Learning* (VOLL). Le niveau de langue de spécialité qui nous intéresse est celui enseigné à l'université, par exemple l'anglais de spécialité enseigné aux étudiants en Master de géographie, qui leur permet de lire et d'écrire des articles scientifiques dans leur domaine.

Il n'existe pas de travaux sur des logiciels dédiés à l'apprentissage des langues de spécialité, ni à la création de ressources pédagogiques spécifiques. Les outils informatisés existants concernant les langues de spécialité se limitent aux logiciels d'aide à la traduction et à des outils d'aide au recensement terminologique (systèmes de gestion de terminologie, extracteurs de termes, bases de données terminologiques, etc.). Par conséquent, notre objectif est de concevoir une plateforme d'aide semi-automatisée à la création de matériels pédagogiques de langues étrangères de spécialité. Cette plateforme doit :

- répondre aux besoins des enseignants (plutôt qu'à ceux des apprenants qui n'en sont pas des utilisateurs directs) ;
- prendre en compte des particularités des langues de spécialité de niveau académique ;
- être utilisable avec plusieurs langues et plusieurs spécialités.

Dans cet article nous décrivons les expérimentations que nous avons mises en place pour la création d'un premier type d'exercices intégrés à la plateforme : des exercices à trous sur le vocabulaire de spécialité. La spécialité a été choisie parmi des thèmes enseignés en cours universitaire de langue étrangère de spécialité : l'anglais de la géographie de l'eau. L'idée mise en œuvre est qu'il est possible de préparer automatiquement ou semi-automatiquement des textes 'authentiques' en langue étrangère de spécialité, choisis et fournis par un enseignant, pour en faire des exercices ou d'autres matériaux pédagogiques pertinents. Les ressources nécessaires pour générer des exercices à trous en langue de spécialité sont :

1 http://hotpot.uvic.ca

- un lexique spécialisé dans le domaine choisi, ce qui se traduit par des listes de termes, éventuellement avec des métadonnées associées aux termes (définitions, exemples, synonymes, etc.).
- Un corpus des textes dans le même domaine, qui sera analysé par la plateforme pour générer des ressources supplémentaires pour la création des exercices.
- Les textes fournis par les enseignants-usagers, qui sont destinés à être transformés en exercices.

Dans la section 2, nous présentons un état de l'art sur les méthodes et outils pour la génération d'exercices pour l'apprentissage des langues, en particulier pour l'apprentissage du vocabulaire et les langues de spécialité. Nous abordons la notion de « contextes riches » qui est au cœur de la problématique de création d'exercices à partir de textes. Dans la section 3 nous décrivons l'expérimentation autour de la génération d'exercices et les résultats. Dans la section 4 nous discutons des pistes de travail pour l'intégration des méthodes proposées dans une plateforme semi-automatique de création de matériels pédagogiques.

2. Etat de l'art

Concevoir des environnements informatiques pour *enseigner* semble un objectif nouveau par rapport à concevoir des environnements informatiques pour *apprendre*, même si le souhait a déjà été exprimé : en 2004, Stéphanie RIOT [RIOT *et al.*, 2004, p. 8] voulait donner à l'enseignant le « rôle majeur » d'utilisateur principal d'un logiciel pédagogique. Plus particulièrement, dans le domaine de la génération automatique d'exercices, le besoin d'outils dédiés aux enseignants a déjà été identifié par Alex BOULTON [BOULTON, 2007, p. 43]. Il propose de promouvoir l'exploitation de corpus en apprentissage des langues par la création de « sites qui complètent des manuels pour un travail autonome ou pour permettre à l'enseignant de créer des activités pertinentes ». Van-Minh PHO [PHO, 2015] signale que les Environnements Informatiques pour l'Apprentissage Humain « doivent fournir des moyens pour assister les enseignants dans leur tâche de génération d'exercices ».
C'est pourtant un logiciel de rédaction manuelle d'exercices qui, dans les revues-conseils, apparaît comme le plus recommandé aux enseignants pour produire des exercices : il s'agit de *HotPatatoes*. Dans son article « Quels logiciels libres pour les professeurs de langues vivantes ? », Laure PESKINE [PESKINE, 2006] répartit l'ensemble des logiciels utilisés par les enseignants en huit catégories : outils bureautiques, outils internet, dictionnaires, traitement des images, traitement du son, lecteur multimédia, création de pages pour publication sur internet, et exerciseurs. Sa dernière catégorie, *exerciseurs*, ne comporte que deux logiciels : *Jclic* et *HotPotatoes* qui ne sont pas spécifiques aux exercices de langue.

2.1. Plateformes de génération d'exercices

A la jonction de l'informatique et de l'apprentissage des langues, le secteur de la génération automatique d'exercices apparaît dispersé quant aux disciplines d'origine des auteurs de publications (professorat, pédagogues, sociétés privées, informaticiens, TAL, etc.). Par ailleurs, les appellations hétéronymes concernant les disciplines et les concepts quasi-communs du secteur sont nombreuses, allant de *logiciel d'édition de contenu pédagogique* à *environnements informatiques pour l'apprentissage humain*. Cette pluralité semble liée au fait que la génération automatique d'exercices est une problématique encore jeune, en pleine expansion.

Il n'existe pas de plateforme dédiée à l'apprentissage de langues de spécialité. Au niveau de l'enseignement universitaire, on trouve les plateformes de génération d'exercices MIRTO et ASKER. Le projet MIRTO [ANTONIADIS *et al.*, 2005] aborde des problématiques didactiques des enseignants de langues, en se centrant sur la création semi-automatique d'exercices de langue générale qui peuvent se succéder sur la plateforme, de manière prédéfinie par un enseignant, pour composer des scénarios qui tiennent compte des réponses des apprenants. La plateforme ASKER [LEFEVRE *et al.*, 2015], elle, est très généraliste : elle sert de support à la création d'exercices dans n'importe quelle matière enseignée à l'université[2]; et, de ce fait, elle n'intègre pas de connaissances relatives aux domaines, lesquelles doivent être apportées par un enseignant.

En ce qui concerne les communautés scientifiques autour de l'enseignement des langues de spécialité, ni les publications du *Groupe d'Étude et de Recherche en Anglais de Spécialité* (GERAS) rassemblées dans la revue *Approches linguistiques des langues spécialisées* (ASP), ni les publications du *Groupe d'Etude et de Recherche en Espagnol de Spécialité* (GERES), ne font mention de logiciels dédiés à l'apprentissage des langues de spécialité du niveau universitaire, ni de logiciels dédiés à l'apprentissage des langues étrangères de spécialité. Les logiciels qui se rapprochent le plus de ces objectifs par la spécialisation de leurs étudiants sont ceux qui aident à l'apprentissage des langues générales au niveau élémentaire dans les filières d'enseignement à l'université [TANO, 2011].

2.2. Génération d'exercices de vocabulaire et vocabulaire de spécialité

Les exercices d'apprentissage des langues peuvent être divisés en 2 catégories [PEREZ-BELTRACHINI *et al.*, 2012] : les exercices basés sur des phrases réelles (*« real life sentences »*, c'est-à-dire les phrases extraites de documents existants), et les exercices basés sur une syntaxe et un vocabulaire limités. Notre travail se situe dans la première catégorie, celle de phrases tirés des documents authentiques, puisque l'apprentissage du terme est aussi important que l'apprentissage de l'environnement dans lequel il est naturellement employé. Il existe plusieurs exerciseurs de ce type [MALAFEEV, 2015], mais aucun ne concerne les langues de spécialité.

Thierry SELVA [SELVA, 2002] décrit plusieurs types d'exercices d'apprentissage de la langue générale dans le cadre du projet E*nvironnement d'apprentissage lexical interactif pour apprenants du français – ALFALEX*. Il décrit notamment les exercices sur les collocations, où un ensemble de phrases est sélectionné dans un corpus pour illustrer les collocations les plus fréquentes. Pour construire les exercices, une partie de la collocation est affichée et l'autre cachée. Le but consiste à compléter la collocation à partir de sa partie affichée et du reste du contexte de la phrase. Le système accepte plusieurs réponses contenant des nuances sémantiques (verbes alternatifs, intensification, etc.) en s'appuyant sur le dictionnaire électronique en ligne *Dictionnaire d'Apprentissage du Français Langue Étrangère ou Seconde - DAFLES*.

Pour les langues de spécialité et le langage technique, Ross CHARNOCK [CHARNOCK, 1999] propose de travailler sur des textes courts (résumés, introductions d'articles, etc.) pour tenir compte des apprenants qui n'ont pas des bases de langue étrangère encore bien établies, tout en supposant qu'ils ont des connaissances adéquates dans la discipline. Il commente l'impossibilité d'un travail efficace sur la langue sans la prise en compte du contexte et des intentions communicatives. Il signale aussi que les textes authentiques de certaines disciplines comportent souvent des archaïsmes linguistiques qui compliquent la tâche des apprenants.

2 Par exemple, elle est actuellement utilisée pour l'enseignement de l'informatique.

Selon MALAFEEV [MALAFEEV, 2015], un système de génération d'exercices basé sur des listes de mots doit prendre en compte des facteurs tels que les majuscules, l'orthographe, la ponctuation, la longueur des mots, la distance entre mots à trouver, le nombre des mots dans le texte, la longueur des mots, etc. Certaines règles établies à partir de ces facteurs permettent de lever des ambiguïtés que l'usage des seuls dictionnaires ne résout pas. D'après Iryna GUREVYCH [GUREVYCH *et al.*, 2009, p. 11], pour l'apprentissage des langues, les paramètres de sélection des mots à remplacer par des blancs dans les exercices à trous peuvent être :

- chaque n-ème mot dans le texte, par exemple n=5 ou n=8 ;
- la fréquence des mots ;
- des mots appartenant à des parties du discours tels que les noms, verbes, adjectifs et adverbes, et dont le sens peut être ciblé ;
- des mots obtenus par un apprentissage automatique basé sur un ensemble de questions saisies (*input questions*) utilisées comme données d'apprentissage.

Pour la création de tests de vocabulaire par les enseignants, Christine COOMBE [COOMBE, 2011] considère le problème du format : le test est valide si les apprenants ont l'expérience du format de présentation du contenu, s'il n'y a pas d'ambiguïté sur comment répondre et comment interpréter les réponses, et si le format a un effet positif sur l'apprentissage, par exemple en aidant la répétition ou l'extension du vocabulaire.

2.3. Les « contextes riches »

Dans un exercice à trous, les termes sur lesquels porte l'exercice sont remplacés par des blancs dans les phrases proposées aux apprenants. Il est donc nécessaire que chaque phrase de l'exercice permette de deviner le terme manquant. Pour la préparation des exercices, pour chaque terme recherché, il faut être en mesure d'identifier automatiquement dans les textes apportés par les enseignants des phrases *riches*, c'est-à-dire des phrases avec un contexte assez riche en informations pour que l'apprenant puisse restituer les termes manquants. Le concept de « contexte riche » est donc important, puisqu'il aide à définir, localiser et prendre en compte les informations contextuelles pertinentes lors de la génération de l'exercice, et ensuite à faciliter la résolution de l'exercice pour un apprenant.

Firas HMIDA *et al.* [HMIDA, 2015] proposent de mettre en œuvre la notion de *Contextes Riches en Connaissances* (CRC) introduite en 2001 par Ingrid MEYER [MEYER, 2001] « pour désigner les contextes qui illustrent des relations entre les termes d'un domaine spécialisé ». Ils proposent l'extraction de 'contextes conceptuels et linguistiques' dans les corpus monolingues spécialisés et dans un corpus scientifique de volcanologie selon deux méthodes :

- la première méthode s'appuie sur la présence du terme à illustrer et l'exploitation d'indices lexicaux pour extraire, grâce à des marqueurs de relations conceptuelles entre termes, des contextes riches en connaissances *conceptuelles* (contextes orientés compréhension) et définir le terme ;
- la seconde méthode s'appuie sur des mesures d'association pour identifier, grâce au repérage de collocations, des contextes riches en connaissances *linguistiques* (contexte orienté usage) et comprendre l'usage du terme.

3. Expérimentation sur la génération automatique des exercices en langues de spécialité

3.1. Préparation du matériel

Notre objectif est de générer des exercices de vocabulaire de spécialité sous forme d'exercices à trous, construits automatiquement à partir d'un texte de spécialité fourni par l'enseignant utilisateur de la plateforme. Le fonctionnement de cet outil est le suivant : la plateforme dispose d'un vocabulaire spécialisé et d'un corpus de textes dans le domaine étudié, que nous appellerons 'corpus support'. L'enseignant apporte un nouveau texte dans le même domaine, que nous appellerons 'texte de référence', à partir duquel seront créés les exercices. L'enseignant choisit les termes à étudier, soit à partir du texte, soit à partir des propositions provenant de la liste du vocabulaire spécialisé. Ces termes sont ensuite recherchés dans le texte de référence pour générer les phrases support de l'exercice, et également dans le corpus, pour fournir à l'apprenant d'autres exemples d'utilisation des termes dans d'autres contextes.

Pour mettre en place notre expérimentation, nous avons choisi le thème d'un cours de Master donné à l'Université de Franche-Comté, *English for Geographers*. Le sous-domaine particulier choisi est celui de *géographie et eau*, en langue anglaise. Les actions à mettre en place sont les suivantes :

- Constituer le corpus support intégrable à la plateforme, pour pouvoir fournir automatiquement des exemples en contexte. Ce corpus sert aussi lors des expérimentations pour vérifier l'adéquation de la liste de vocabulaire spécialisé à des textes du domaine.
- Établir une liste de termes de spécialité intégrable à la plateforme.
- Constituer un ensemble de textes de référence pour tester la plateforme.

Corpus support
Le corpus support est constitué de 44 textes en anglais provenant de Wikipédia, édités entre 2013 et 2015, soit un total de 199448 mots. L'utilisation de l'encyclopédie en ligne Wikipédia est due au fait qu'elle propose un large choix de textes spécialisés en accès libre, et permet d'envisager une automatisation du choix des textes dans l'avenir. Le corpus a été nettoyé manuellement et converti en format TXT.

Liste de vocabulaire de spécialité
Il s'agit de constituer une liste des termes d'intérêt de la langue de spécialité, sans être spécialiste du domaine, ce qui est le cas de l'enseignant, futur utilisateur de la plateforme. Cette liste constituera le vocabulaire de spécialité intégré à la plateforme.

Pour une première expérimentation nous avons choisi d'utiliser deux listes terminologiques déjà existantes : *International glossary of hydrology* [OMM, 2012], qui contient 2059 termes, et le *lexique anglais-français* du *Dictionnaire encyclopédique des sciences de l'eau* [RAMADE, 1998], qui contient 1645 termes.

Ensemble de textes de référence
Les textes de référence sont les textes fournis par l'enseignant. C'est dans les phrases de ces textes que des termes de spécialité vont être choisis pour être remplacés par des blancs et présentés aux apprenants sous la forme des exercices à trous.

Il importe de noter que les phrases extraites du corpus support (qui sont différentes de celles du texte fourni par l'enseignant) serviront, elles, d'indices de contexte supplémentaire pour aider l'apprenant à deviner les termes remplacés par des blancs dans le texte de référence.

L'ensemble de textes de référence est constitué de 20 textes sur la géographie de l'eau publiés entre 2009 et 2016, dont 10 textes scientifiques (issus du journal *Water Research* de l'*International Water Association - IWA*, et de l'organisation d'éducation environnementale *Field Studies Council - FSC*) et 10 textes journalistiques spécialisés (issus de la *National Geographic Society*, du magazine en ligne *ScienceDaily*, de la *Royal Geographical Society* et de la *BBC*). Ces textes ont été nettoyés manuellement et convertis en format TXT.

3.2. Expérimentations et résultats

Listes du vocabulaire de spécialité

Pour obtenir la liste de vocabulaire spécialisé à intégrer dans la plateforme, nous avons combiné deux listes terminologiques cités précédemment : *International glossary of hydrology (A)* et *Dictionnaire encyclopédique des sciences de l'eau* (*B*). Nous avons identifié les termes communs entre ces deux listes (leur intersection A ∩ B), l'objectif étant d'évaluer les différences entre les deux listes. Nous avons également constitué la liste de termes appartenant à la liste A ou B (leur union).

Nous avons projeté chacune de ces listes sur le corpus support. Cette tâche a été effectuée par un script qui permet d'identifier toutes les occurrences des termes dans les textes dans leurs formes au singulier et au pluriel. Le tableau 1 présente les résultats.

Liste	Nombre de termes	Nombre d'occurrences dans le corpus	Nombre de termes (uniques) dans le corpus	Pourcentage des termes qui apparaissent dans le corpus
A	2 059	11 268	565	27,44 %
B	1 645	19 748	543	33,01 %
A ∩ B	202	8 032	155	76,73 %
A ∪ B	3 502	21 779	952	27,18 %

Tableau 1 : Projection des termes des listes sur le corpus support

Nous constatons que les listes *A* et *B* ont peu de termes en commun : 202 termes, ce qui constitue moins de 10 % pour *A* et moins de 13 % pour *B*. Nous avons constaté que les deux listes ne contiennent pas les mêmes classes sémantiques de termes dans les mêmes proportions : par exemple, la liste *B* contient plus de noms d'espèces vivantes que la liste *A*. Malgré ces disparités, les 202 termes communs nous ont permis de constituer une catégorie expérimentale de termes centraux de la spécialité géographie-eau pour la sélection de termes des exercices à trous. Ces 202 termes communs sont très bien représentés dans le corpus : 155 parmi eux ont des occurrences dans le corpus.

Notons également que les listes *A* et *B* sont tout à fait comparables, à la fois en nombre de termes et en proportion de termes reconnus dans le corpus. De ce fait, nous avons choisi d'intégrer l'union de ces deux listes (A ∪ B) à la plateforme.

Afin d'évaluer la pertinence de la liste A ∪ B pour la création d'exercices à partir de textes du domaine géographie-eau, nous l'avons projeté sur le corpus de textes de référence. Le tableau 2 présente les résultats. La dernière colonne donne le pourcentage des termes de la liste A ∪ B qui ont des occurrences dans le texte.

Texte	Nombre de mots	Nombre d'occurrences	Nombre de termes	Pourcentage
Textes journalistiques (10 textes)				
1	815	120	52	6,38 %
2	2 177	306	82	3,77 %
3	1 229	228	71	5,78 %
4	1 215	151	52	4,28 %
5	915	141	54	5,90 %
6	612	100	29	4,74 %
7	1 259	261	77	6,12 %
8	4 581	85	37	0,81 %
9	491	115	46	9,37 %
10	1 833	285	78	4,26 %
Textes scientifiques (10 textes)				
11	6 694	1 040	140	2,09 %
12	10 660	1 433	109	1,02 %
13	15 051	2 971	263	1,75 %
14	10 461	2 168	151	1,44 %
15	8 833	1 947	142	1,61 %
16	14 731	2 501	202	1,37 %
17	3 561	738	145	4,07 %
18	2 312	402	84	3,63 %
19	2 901	610	111	3,83 %
20	2 494	436	87	3,49 %

Tableau 2 : Projection des termes de la liste A ∪ B sur le corpus de référence

Reconnaissance de phrases significatives

Par *phrase significative* nous entendons une phrase qui, premièrement, inclut le terme que l'on veut effacer par un blanc, et, deuxièmement, permet de deviner ce terme, grâce à des caractéristiques contextuelles reconnaissables que nous appellerons 'indices'. Nous avons procédé au repérage manuel de phrases significatives du corpus support parmi celles contenant les termes projetés, pour établir une liste des indices, dont une partie sont présentés dans le tableau 3.

Nous associons à chaque indice détecté un poids, qui exprime son degré d'informativité ou richesse de contexte. Nous l'appellerons *poids d'indice*.

Dans la phrase, le poids du terme que l'on veut effacer par un blanc est obtenu par l'addition des poids de tous les indices qui le concernent dans la phrase et dans le contexte proche. Nous proposons de choisir les phrases significatives en tenant compte du poids du terme le plus élevé (ou plusieurs termes de poids élevé).

Caractéristiques contextuelles (indices)	**Poids**
Le terme est un hapax : il n'y a qu'une seule occurrence du terme dans le texte. Exemple (1) : 'hydrological models'.	+2
Indice de description ou définition dans la phrase, après le terme. Exemple (1) : 'is the branch of', 'deals with'.	+2
Présence d'autres termes projetés dans la phrase, quelque soit leur nombre. Exemple (3) : 'source, 'sources'.	+1
Présence d'autres termes projetés dans la phrase suivante, quelque soit leur nombre. Exemple (2) : 'processes'.	+0,5
Indice d'explication dans la phrase, après le terme : 'for example','that means',...	+1,5
Indice 'faible' d'extension de la phrase, après le terme (ne compter qu'une fois chacun de ces indices, quel que soit leur nombre dans la phrase). Exemple (3) : '(', 'or', 'and', ','.	+0,5
Indice 'fort' d'extension de la phrase, après le terme : 'however', 'also', 'but',...	+1
Indice d'anaphore dans la phrase, après le terme. Exemple (1) : 'which'.	+1,75
Indice d'anaphore dans la phrase, avant le terme.	-1
Indice d'anaphore dans le 1er syntagme de la phrase suivante : 'it', 'that',… Exemple (2) : 'They'.	+1
Le terme est immédiatement suivi du verbe être, à la 3ème personne de son nombre (singulier ou du pluriel) : 'is', 'are'. Exemple (2) : 'are'.	+1,5

Tableau 3 : Exemples de caractéristiques contextuelles

Voici des exemples de pondérations correspondant à des indices du tableau 3, pour évaluer la richesse des phrases ((a) phrases d'origine, (b) mêmes phrases avec pondération, et (c) calcul du *poids d'indice* par addition des pondérations du contexte pour le terme) :

(1) Poids d'indice fort : 9,25 pour le contexte du terme 'hydrography' :

(1.a) « hydrography is the branch of applied sciences which deals with the measurement and description of the physical features of oceans, seas, coastal areas, lakes and rivers,... »

(1.b) « **[hydrography]** '**is the branch of**'**(2)** applied sciences '**which**'**(1,75)** '**deals with**'**(2)** the measurement **and(0,5)** '**description of**'**(2)** the physical features of oceans, **[seas](1/4)**, coastal **[areas](1/4)**, **[lakes](1/4)** and **[rivers](1/4)**,... ».

(1.c) **[hydrography]** = 2 + 1,75 + 2 + 0,5 + + 2 + 1 **= 9,25**

(2) Poids d'indice moyen : 4,5 pour le contexte du terme 'hydrological models' :

(2.a) « hydrological models are simplified, conceptual representations of a part of the hydrologic cycle. They are primarily used for hydrological prediction and for understanding hydrological processes ».

(2.b) « **[hydrological models] 'are'(1,5)** simplified **','(0,5)** conceptual representations of a part of the hydrologic **[cycle](1)**. '**They**'**(1)** are primarily used for hydrological prediction and for understanding hydrological **[processes](0,5)** ».

(2.c) **[hydrological models]** = 1,5 + 0,5 + 1 + 1 + 0,5 **= 4,5**

(3) Poids d'indice faible : 3 pour le contexte du terme 'river' :

(3.a) « A river begins at a source (or more often several sources) and ends at a mouth, following a path called a course »

(3.b) « A **[river]** begins at a **[source](1/2) '('(0,5) 'or'(0,5)** more often several **[sources](1/2) 'and'(0,5)** ends at a mouth **','(0,5)** following a path called a course ».

(3.c) **[river]** = 1/2 + 0,5 + 0,5 + 1/2 + 0,5 + 0,5 **= 3**

En résumé, nous proposons d'appliquer aux textes (ceux du corpus support et ceux fournis par l'enseignant-usager de la plateforme) une pondération des indices comme celle décrite ci-dessus, afin de détecter 1) les phrases significatives dans le corpus support (elles doivent servir d'exemples ajoutés dans l'exercice à trous), 2) les termes à remplacer par des blancs dans le texte de référence fourni par l'enseignant-usager.

4. Conclusion

Afin de développer une plateforme pour les enseignants de langues étrangères de spécialité pour la préparation de matériels didactiques, nous avons posé les bases d'un générateur automatique d'exercices à trous en langues de spécialité. Nous avons tout d'abord conceptualisé le fonctionnement d'un tel exerciseur en nous basant sur les exerciseurs qui existent déjà pour l'apprentissage des langues étrangères (non spécialisées). Nous avons sélectionné des matériels de langues de spécialité intégrables à la plateforme selon des procédures réutilisables pour d'autres langues et spécialités, et nous avons effectué une expérimentation de méthodes de sélection des termes et des phrases dans le corpus pour produire automatiquement des exercices à trous. Nous avons mis en place des indices permettant d'exprimer et calculer la valeur informative d'une phrase, pour qu'elle constitue un contexte significatif pour la recherche d'un terme.

Afin de générer des exercices à choix multiples, nous devons considérer les suggestions des réponses qui pourraient être données à un apprenant cherchant à deviner un terme dans une phrase à trou. Pour que ces suggestions soient cohérentes, nous avons effectuée une première catégorisation sémantique pour les termes du domaine « géographie-eau ». Un terme peut appartenir à plusieurs catégories à la fois. Quelques exemples de catégories sont présentés dans le tableau 4.

Catégorie	Termes
Lieu	'river bed', 'river bank', 'meander', 'mouth', 'estuary', 'delta', 'cliff face', 'coastline', 'source', 'bridge', 'harbour',...
Etat	'liquid', 'humid', 'temperate', 'tropical', 'cold', 'dry', 'hot', 'warm', 'wet', 'polar',...
Phénomène	'evaporation', 'water level', 'condensation', 'flooding', 'confluenc'e, 'melting', 'freezing', 'erosion', 'deposition', 'attrition', 'soil erosion', 'deforestation', 'flooding', 'monsoon', 'erosion', 'abrasion', 'flow down', 'storm', 'thunderstorm', 'drought',...
Concept	geography, location, water level, weather, confluence,...
Objet naturel	'water vapour', 'glacier', 'lake', 'source', 'cloud', 'coast', 'wave', 'cliff', 'biomes', 'landscapes', 'tundra', 'channel', 'desert', 'cloud', 'ecosystem', 'environment', 'fog', 'ice', 'population', 'rain', 'sea', 'river', 'snow', 'sun', 'water', 'wind', 'sediment', 'storm', 'thunderstorm',...
Objet technique	'bridge', 'harbour', 'dam', 'aqueduct', 'map',...
Vivant	'deforestation', 'biomes', 'tundra', 'ecosystem', 'environment', 'population',...

Tableau 4 : Exemples de catégories sémantiques spécifiques au domaine de spécialité

Les résultats de ces premières expérimentations montrent la faisabilité de la génération d'exercices à trous à partir de listes de vocabulaire et textes de référence. Dans le futur, nous allons développer l'outil de génération d'exercices à trous, et nous allons construire, sur la base de ce type d'exercices, un prototype opérationnel de la plateforme de génération automatique d'exercices de langues de spécialités et autres matériels pour les enseignants de langues de spécialité.

Références

ANTONIADIS Georges, ECHINARD S., KRAIF Olivier, LEBARBE T., PONTON Claude (2005), « Modelisation de l'integration de ressources TAL pour l'apprentissage des langues : la plateforme MIRTO », in *Alsic*, vol. 8, n° 2 spécial Atala.

BOULTON Alex (2007). « Esprit de corpus: Promouvoir l'exploitation de corpus en apprentissage des langues ». *Texte et Corpus*, n° 3, 37-46.

CHARNOCK Ross (1999). « Les langues de spécialité et le langage technique : considérations didactiques », *ASp*, n° *23-26*, 281-302.

COOMBE Christine A. (2011). « Assessing Vocabulary in the Language Classroom » in Anderson & Sheehan (Eds) « Focus on Vocabulary: Emerging Theory and Practice for Adult Language Learners », 111-124. HCT Press, Abu Dhabi.

GUREVYCH Iryna, BERNHARD D. et BURCHARDT A. (2009). « Tutorial Notes - Educational Natural Language Processing », *AIED 2009*, Brighton. Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt, Allemagne.

HMIDA Firas, MORIN E. et DAILLE B. (2015). « Extraction de Contextes Riches en Connaissances en corpus spécialisés » in Actes de la *22ème conférence sur le Traitement Automatique des Langues Naturelles*, 425-431, Caen.

L'HOMME Marie-Claude (1990). « Y a-t-il une langue de spécialité ? Points de vue pratique et théorique » in revue *Langues et linguistique*, numéro spécial Journées de linguistique, 2011, 26-33. Centre international de recherche en aménagement linguistique, Québec.

LEFEVRE Marie, GUIN N., CABLÉE B. et BUFFA B. (2015). « ASKER : un outil auteur pour la création d'exercices d'auto-évaluation ». Atelier Evaluation des Apprentissages et Environnements Informatiques – EAEI, *Conféerence EIAH 2015*, Agadir, Maroc.

MEYER Ingrid (2001). « Extracting knowledge-rich contexts for terminography - A conceptual and methodological framework » in B. DIDIER, J. CHRISTIAN et M.-C. L'HOMME (Eds), *Recent Advances in Computational Terminology*, 279–302. Cité par : [HMIDA et al., 2015].

MALAFEEV Alexey Yurievich, (2015). « Exercise Maker: Automatic Language Exercise Generation » in *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2015)* n° 14(21), 441-452. Russian State University for the Humanitie. National Research University Higher School of Economics, Nizhny Novgorod, Russie.

Organisation météorologique mondiale - OMM et Organisation des Nations unies pour l'éducation, la science et la culture - UNESCO (2012). « Glossaire international d'hydrologie », publication OMM n° 385, 3ème édition (ouvrage quadrilingue anglais, français, russe, espagnol). Genève, Suisse : OMM. URL :
www.wmo.int/pages/prog/hwrp/publications/international_glossary/385_IGH_2012.pdf

PEREZ-BELTRACHINI Laura, GARDENT C. et KRUSZEWSKI G. (2012). « Generating Grammar Exercises » in The 7th Workshop on Innovative Use of NLP for Building Educational Applications, *NAACL- HLT Worskhop 2012*, 147-157, Montréal.

PESKINE Laure (2006). « Quels logiciels libres pour les professeurs de langues vivantes ? » in revue *Langues Modernes n° 1*. Paris : Association des Professeurs de Langues Vivantes - APLV.

PHO Van-Minh (2015). « Génération automatique de questionnaires à choix multiples pédagogiques : évaluation de l'homogénéité des options ». Thèse de doctorat, LIMSI-CNRS, Université Paris Sud - Paris XI.

RAMADE François (1998). « Lexique anglais-français » in *Dictionnaire encyclopédique des sciences de l'eau : biogéochimie et écologie des eaux continentales et littorales*, 715-735. Paris : Édiscience International.

RIOT Stéphanie, GUIN N., JEAN-DAUBIAS S. (2004). « Assistance à l'enseignant dans le cadre de l'EIAH AMBRE : conception d'un générateur de problèmes », rapport de recherche LIRIS (stage de DEA Informatique et PFE INSA), LIRIS - CNRS.

SELVA Thierry (2002). « Génération automatique d'exercices contextuels de vocabulaire » in *Actes de TALN 2002*, 185-194, Nancy.

TANO Marcelo, (2011). « L'utilisation de plateformes en ligne dans l'enseignement apprentissage de l'Espagnol pour Objectifs Spécifiques » in « Innovations didactiques dans l'enseignement apprentissage de l'espagnol de spécialité grâce aux ressources technologiques », *Les cahiers du GERES (Groupe d'Étude et de Recherche en Espagnol de Spécialité), nº 4*, 77-102. Montpellier.

Origines des erreurs en Traduction Spécialisée : différentiation textométrique grâce aux corpus de textes cibles annotés

Natalie Kübler[1], Maria Zimina[1], Serge Fleury[2]
(1) CLILLAC-ARP EA 3967, Université Paris Diderot-Paris 7, France
(2) CLESTHIA EA 7345, Sorbonne Nouvelle-Paris 3, Paris, France
nkubler@eila.univ-paris-diderot.fr, mzimina@eila.univ-paris-diderot.fr,
serge.fleury@univ-paris3.fr

RESUME

L'étude présente une analyse quantitative de traductions annotées selon une typologie d'erreurs, en vue de l'amélioration des méthodologies d'enseignement de la traduction spécialisée (TS). Les productions annotées sont alignées avec les textes originaux au niveau de la phrase. Les spécificités morpho-syntaxiques sur les contextes sources regroupés par types d'erreurs de traduction permettent de récolter des indices sur les éléments complexes des discours spécialisés qui génèrent des erreurs lors du processus de transfert du sens. La visualisation contextuelle de ces éléments via la Lecture Textométrique Différentielle (LTD) ouvre des perspectives pour la conception des modules de prise en charge des difficultés caractéristiques des apprenants en TS.

ABSTRACT

Origins of errors in Specialized Translation: textometric differentiation through annotated target text corpora

This study focuses on the quantitative analysis of translations annotated according to an error typology. The purpose of the study is to improve current methodologies of Specialized Translation (SP) teaching. We conduct a quantitative analysis of annotated translations aligned with their original texts. Characteristic elements are computed on morpho-syntactic level of analysis in source contexts grouped according to translation errors in the target text. They reveal specific linguistic elements that are complex in terms of meaning transfer in specialized languages. Contextual visualization of these results through Differential Textometric Browsing (DTB) opens up new horizons for the development of language modules based on the types of difficulties that learners face in SP.

MOTS-CLES : Annotation d'erreurs, corpus alignés, enseignement de la Traduction Spécialisée (TS), textométrie, Lecture Textométrique Différentielle (LTD), méthode des spécificités, visualisation.

KEYWORDS: Aligned corpora, characteristic elements computation, Differential Textometric Reading (DTR), annotation, Specialized Translation teaching, textometric analysis, visualisation.

Natalie Kübler, Maria Zimina, Serge Fleury

1 Introduction

1.1 Contexte de l'étude

Récemment, les approches centrées sur le processus de traduction et sur l'évaluation des méthodes d'enseignement ont fait l'objet de plusieurs travaux de recherche (Bowker, Bennison, 2003 ; Pearson, 2003 ; Castagnoli et al., 2011 ; Loock et al., 2014 ; Frankenberg-Garcia, 2015). En Traduction Spécialisée (TS), beaucoup de chantiers restent encore à explorer pour mieux cerner les difficultés de transfert de sens mobilisant des compétences linguistiques et cognitives très variées face à la complexité des textes en langues de spécialité (Kübler, 2011 ; Froeliger, 2013). Dans ce contexte, la recherche expérimentale à base de corpus offre la possibilité d'observer de manière systématique les stratégies employées par les traducteurs qui peuvent passer inaperçues avec d'autres approches (Granger et al., 2002 ; Frérot, 2010).

Notre travail s'appuie sur une méthodologie d'enseignement de la traduction spécialisée qui permet une identification systématique des problèmes de traduction par le biais d'une analyse textométrique de corpus de traduction annotés selon la typologie d'erreurs MeLLANGE (Castagnoli et al., 2011), avec l'aide du logiciel *Le Trameur* (Fleury, Zimina, 2014). Ce cadre méthodologique est mis en place depuis septembre 2013 à l'université Paris Diderot dans le cadre du Master 1 Industrie des langues et Traduction spécialisée (ILTS).[1]

Dans un premier temps, la méthodologie employée nous a permis de distinguer les catégories d'erreurs les plus saillantes et les schémas morphosyntaxiques qui les caractérisent en contexte, à savoir les problèmes touchant les termes composés et complexes, mais également les collocations, les prépositions, les verbes, etc. Les analyses faites sur les productions des apprenants de ces deux dernières années (2013-2015) nous ont aussi amenés à intégrer à l'enseignement de la TS une approche plus ciblée des problèmes posés par la traduction des termes spécialisés et des GN complexes (Kübler et al., 2016).

Dans cette nouvelle étude, nous nous intéressons aux origines des erreurs en traduction spécialisée par profilage des contextes sources à partir des annotations des erreurs dans les productions en langue cible. Nous pensons que les corpus de traductions alignées avec les textes sources peuvent aider à récolter des indices sur les éléments complexes des discours spécialisés qui sont à l'origine des erreurs récurrentes. La découverte de ce type d'indices peut ensuite alimenter la réflexion sur le développement des modules spécifiques qui ciblent les problèmes récurrents des apprenants.

1.2 Objectifs

L'objectif de cette expérimentation est la création de cours spécifiquement adaptés aux problèmes identifiés dans les textes sources. Ce travail vise à contribuer au développement de la méthodologie d'enseignement de la TS qui mêle par ailleurs des compétences transversales à plusieurs niveaux. Pour les étudiants, il s'agit de découvrir et de se former à l'approche de la traduction basée sur le

[1] http://www.eila.univ-paris-diderot.fr/formations-pro/masterpro/ilts/index

corpus, au travail sur les discours spécialisés en collaboration étroite avec les experts du domaine, et à l'analyse terminologique et notionnelle préalable au processus de traduction.

La prise en compte de l'alignement des textes originaux et des traductions annotées selon la typologie d'erreurs MeLLANGE vise à élaborer des propositions méthodologiques à base d'exemples concrets à l'origine des erreurs de traduction. Cet ancrage dans le texte source constitue le trait caractéristique de ce volet de l'étude.

Notre approche est exploratoire : les corpus sont exploités pour détecter des éléments complexes dans les contextes sources à partir des profils d'erreurs repérées dans la traduction. Sur ce plan, on peut constater des similitudes entre les objectifs de cette étude et des travaux sur le profilage d'erreurs de la traduction automatique (Kübler et al., 2013 ; Wisniewski et al., 2014). Dans les deux cas, il s'agit de recueillir des informations sur les origines des erreurs et d'en tenir compte dans les nouvelles productions.

2 Corpus aligné et annoté ER-TRAD-SP1 (anglais-français)

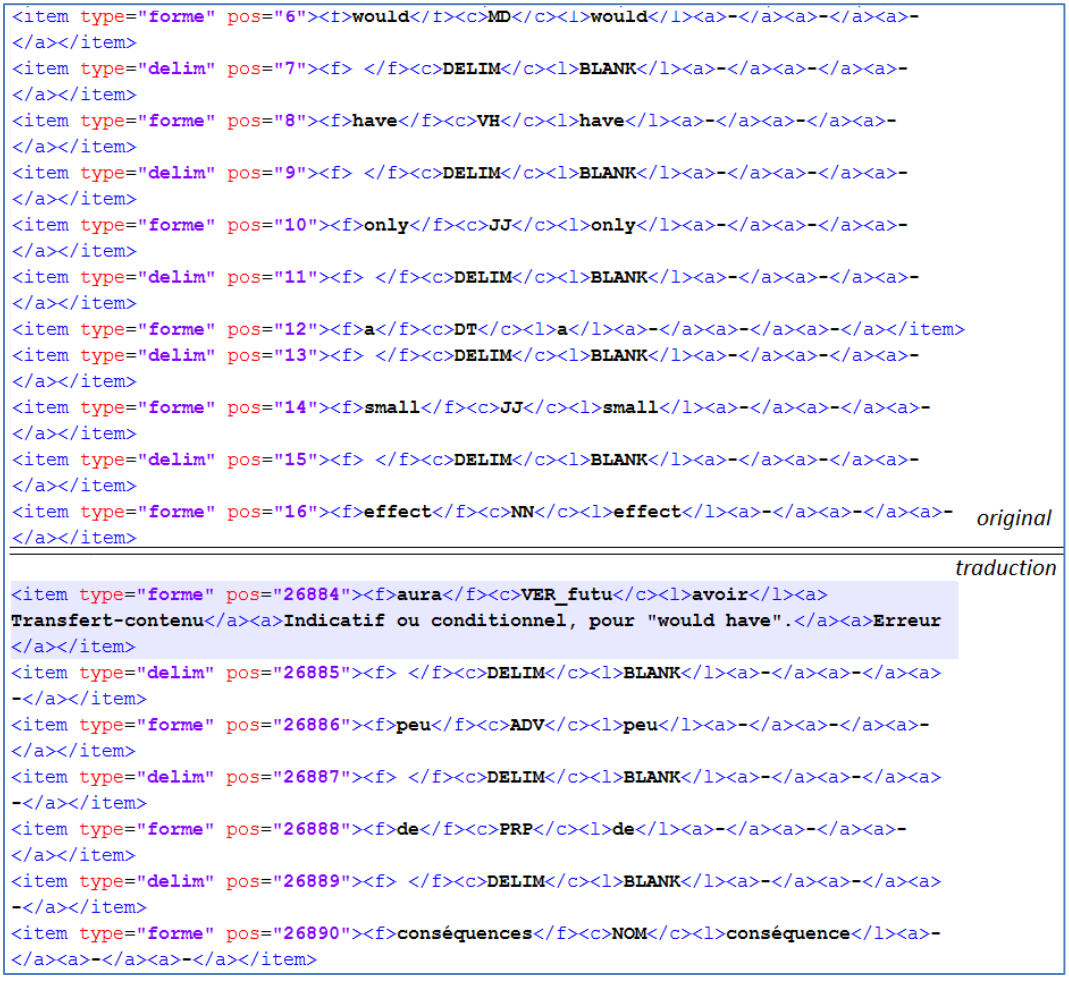
La présente étude exploite les traductions de l'anglais vers le français réalisées par les étudiants M1 en 2014-2015. Il s'agit de 55 extraits de 14 articles scientifiques en Sciences de la Terre (37 324 occurrences de formes graphiques au total). Ce corpus est subdivisé en deux sous-corpus : *ER-TRAD-SP1* (15 311 occurrences de formes graphiques) constitué de traductions réalisées sans accès aux corpus, et *ER-TRAD-SP2* (22 013 occurrences de formes graphiques) constitué de traductions réalisées avec l'aide du corpus. Les traductions portent sur des extraits d'articles scientifiques avec une très haute densité terminologique et le registre de langue caractéristique du discours scientifique. Les problèmes de traduction qu'affrontent les apprentis traducteurs sont nombreux et variés. L'annotation des erreurs selon la typologie MeLLANGE permet toutefois une catégorisation fine des différents problèmes rencontrés dans ce type de discours. Au total, le sous-corpus *ER-TRAD-SP1* compte 886 annotations ; le sous-corpus *ER-TRAD-SP2* compte 893 annotations (Kübler et al., 2016).

Actuellement, le corpus *ER-TRAD-SP1* (traductions sans accès au corpus) a été aligné au niveau de la phrase avec les textes originaux. Ce travail a été réalisé à l'aide d'une série de scripts en s'appuyant sur les alignements phrastiques initialement proposés par les étudiants au cours de la traduction. Nous avons également fait appel aux fonctions du programme *MkAlign*[2] pour la vérification et synchronisation finale de l'alignement. Chaque volet du corpus a été étiqueté par *TreeTagger*[3] intégré dans *Le Trameur*[4] et converti en une base textométrique au format XML dans laquelle l'annotation des erreurs de traduction est renseignée pour le volet français.

[2] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[3] http://www.tal.univ-paris3.fr/mkAlign/
[4] http://www.tal.univ-paris3.fr/trameur/

La Figure 1 montre deux segments issus de l'alignement des phrases dans la base textométrique *ER-TRAD-SP1* au format lu par *Le Trameur*. Cette base comporte 6 niveaux d'annotation. Pour chaque *item* dénombré (type forme ou délimiteur), on retrouve :

1. <f> sa forme graphique </f>
2. <c> sa catégorie morpho-syntaxique </c>
3. <l> son lemme </l>
4. <a> son type d'erreur MeLLANGE (ou son absence)
5. <a> le commentaire éventuel de l'annotateur
6. <a> l'indication sur la présence d'erreur tous types confondus (ou son absence)



```
<item type="forme" pos="6"><f>would</f><c>MD</c><l>would</l><a>-</a><a>-</a><a>-
</a></item>
<item type="delim" pos="7"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a><a>-
</a></item>
<item type="forme" pos="8"><f>have</f><c>VH</c><l>have</l><a>-</a><a>-</a><a>-
</a></item>
<item type="delim" pos="9"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a><a>-
</a></item>
<item type="forme" pos="10"><f>only</f><c>JJ</c><l>only</l><a>-</a><a>-</a><a>-
</a></item>
<item type="delim" pos="11"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a><a>-
</a></item>
<item type="forme" pos="12"><f>a</f><c>DT</c><l>a</l><a>-</a><a>-</a><a>-</a></item>
<item type="delim" pos="13"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a><a>-
</a></item>
<item type="forme" pos="14"><f>small</f><c>JJ</c><l>small</l><a>-</a><a>-</a><a>-
</a></item>
<item type="delim" pos="15"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a><a>-
</a></item>
<item type="forme" pos="16"><f>effect</f><c>NN</c><l>effect</l><a>-</a><a>-</a><a>-
</a></item>
```

original

traduction

```
<item type="forme" pos="26884"><f>aura</f><c>VER_futu</c><l>avoir</l><a>
Transfert-contenu</a><a>Indicatif ou conditionnel, pour "would have".</a><a>Erreur
</a></item>
<item type="delim" pos="26885"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a><a>
-</a></item>
<item type="forme" pos="26886"><f>peu</f><c>ADV</c><l>peu</l><a>-</a><a>-</a><a>-
</a></item>
<item type="delim" pos="26887"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a><a>
-</a></item>
<item type="forme" pos="26888"><f>de</f><c>PRP</c><l>de</l><a>-</a><a>-</a><a>-
</a></item>
<item type="delim" pos="26889"><f> </f><c>DELIM</c><l>BLANK</l><a>-</a><a>-</a><a>
-</a></item>
<item type="forme" pos="26890"><f>conséquences</f><c>NOM</c><l>conséquence</l><a>-
</a><a>-</a><a>-</a></item>
```

FIGURE 1 : Base textométrique alignée ER-TRAD-SP1 (extraits)

3 Méthodes : analyse des origines d'erreurs en TS par la différentiation textométrique

3.1 Diagnostics de spécificités sur contextes alignés

La *textométrie multilingue* (Fleury, Zimina, 2014 ; Zimina, Fleury, 2015) propose des méthodes quantitatives adaptées à l'observation des variations de fréquence d'unités textuelles (formes, lemmes, catégories, etc.) en contextes alignés. En suivant cette approche, nous mobilisons les profils d'erreurs en contextes cibles pour amorcer l'analyse quantitative des contextes sources correspondants. Les contextes d'erreurs de traduction (phrases alignées) sont analysés en termes de *spécificités* avec *Le Trameur*. Cette méthode permet de mesurer les variations de la fréquence dans un corpus découpé en parties (Lebart, Salem, 1994). Dans notre cas, le contexte d'erreur correspond à la phrase et la comparaison s'effectue entre deux parties : les phrases avec et sans erreurs de traduction en langue cible. Les empans d'erreurs (calculés en nombre d'occurrences de formes graphiques) varient selon les types d'erreurs.[5]

La *méthode des spécificités* (Lafon, 1984) met en évidence pour chaque unité de décompte les parties de corpus dans lesquelles l'unité possède de nombreuses occurrences (spécificités positives) ainsi que celles où son effectif est au contraire anormalement faible (*spécificités négatives*). On calcule le diagnostic de spécificité relatif à l'effectif constaté à base des paramètres suivants : k_{ij} - sous-fréquence de l'unité dans la partie, F_i - fréquence de l'unité dans l'ensemble du corpus, T_j - nombre des unités dans la partie, T - nombre total des unités du corpus. Un calcul probabiliste permet de porter un jugement sur l'effectif analysé (k_{ij}) compte tenu des trois autres nombres (F_i, T_j, T). Si l'effectif k_{ij} se situe dans les limites de ce que le calcul permettait d'espérer, la répartition constatée est considérée « banale ». Si ce n'est pas le cas, on calcule un *indice de spécificité* de l'unité. Le diagnostic est fourni sous la forme ±*xx* où le signe (+ ou -) indique un sur-emploi ou un sous-emploi de l'unité dans la ou les partie(s) sélectionnée(s) par rapport à l'ensemble du corpus ; *xx* est un indice de spécificité qui est d'autant plus élevé que la sous-fréquence analysée s'écarte d'une répartition « neutre » qui est sous-jacente au modèle des *spécificités*.[6]

Sur la Figure 2, les diagnostics de *spécificités* sont présentés sous forme de listes de catégories morpho-syntaxiques caractéristiques des contextes d'erreurs répertoriées dans la traduction. Pour chaque erreur, les catégories surreprésentées dans les contextes sources sont triées par la valeur d'*indice de spécificité* (du plus haut vers le plus bas). Seuls les profils sources d'erreurs fréquentes sont représentés (lorsque la zone de couverture de l'annotation cible est supérieure à 50 occurrences de formes graphiques annotées). Cette liste des *spécificités* sert de point d'entrée pour explorer les profils d'erreurs à l'aide des fonctionnalités disponibles dans *Le Trameur* : analyse des sections alignées, concordances, graphiques de ventilation, cartographie différentielle sur corpus parallèle, etc. (Zimina, Fleury, 2015).

[5] Sur le calcul des longueurs moyennes des empans d'erreurs, consulter Kübler et al. (2016).

[6] Le modèle probabiliste utilisé ici pour évaluer la répartition est le modèle hypergéométrique, cf. Lafon (1984, pp. 54-68). Sur la pratique du calcul des *spécificités* on consultera également Lebart, Salem (1994, pp. 172-176).

Erreur MeLLANGE (Nb. occ. formes annotées)	Eléments caractéristiques des contextes sources : *spécificités positives* sur catégories *TreeTagger* (extraits d'articles en Sciences de la Terre)	Indice de spécificité (seuil de 10)
Distorsion (395 occ. formes annotées)	Verb, past participle (*considered*, *derived, inspected*) Modal verb (*can*, *could, may, must, should, would*)	+3 +2
Formulation maladroite (366 occ. formes annotées)	Complementizer *that* Verb, gerund/participle (*melting, bearing*) *Wh*-abverb (*when, where*) Verb *be*, present non-3rd p. (*are*)	+3 +3 +2 +2
Trop littérale (286 occ. formes annotées)	Verb, past participle (*compared, formed, shown*) List marker (*1, 2, b, d*) Modal verb (*might, should, would*) Adverb, comparative (*more*) Adverb (*essentially, respectively, slightly*) Verb *be*, base form Verb, base form (*generate, induce, melt*)	+3 +3 +3 +2 +2 +2 +2
Terme traduit par non terme (150 occ. formes annotées)	Noun singular or massive (*mantle*, *olivine, subduction*) Verb *be*, past (*was, were*) Verb, past participle (*analysed*, *derived, used*) Verb *be*, present non-3rd p. (*are*)	+3 +3 +2 +2
Choix incorrect (128 occ. formes annotées)	*Wh*-determiner (*which*) Modal verb (*can*, *may, should*) Determiner (*the*) Verb, past tense (*observed*, *increased*) Verb, present, non-3rd p. (*conclude, suggest*)	+3 +2 +2 +2 +2
Omission (98 occ. formes annotées)	Verb *have* present, non-3rd p. (*have*) Verb, past participle (*formed*, *recorded*, *correlated*) Verb *be*, present 3rd p. sing (*is*)	+3 +2 +2
Collocation incorrecte (93 occ. formes annotées)	Verb *be*, past (*was, were*) Cardinal number (*300, 500, two*) Verb *have* present, non-3rd p. (*have*) Verb, past participle (*reported, based*)	+5 +3 +2 +2
Syntaxe (70 occ. formes annotées)	Modal verb (*can, cannot, may*) Noun singular or massive (*buoyancy*, *crust*, *pressure*) Noun plural (*data, lavas, measurements*) Verb, present, non-3rd p. (*affect, conclude, denote*)	+3 +3 +2 +2

TABLE 1: Profilage des contextes sources d'erreurs de traduction

Origines des erreurs en TS : différentiation textométrique grâce aux corpus de textes cibles annotés

On remarque que les profils des contextes sources de certaines erreurs sont proches. Dans ce cas, il s'agit le plus souvent des erreurs en relation de cooccurrence. Par exemple, les erreurs type « Distorsion » (77 contextes) et celles qui relèvent des problèmes de « Formulation maladroite » (72 contextes) partagent 15 contextes (phrases), par exemple :

> *(Original) The current melting points are up to 1000 K higher than the melting points obtained by the observation of surface motion of a laser-heated sample (8).§*
>
> *(Traduction) : Les points de fusion actuels ont une* ***température_supérieure, _jusqu'à_1000_K, _aux_points_de_fusion_observés****[Formulation-maladroite_LA-ST-AW] sur* ***la_surface_en_mouvement****[Distorsion_TR-DI] de l'échantillon chauffé par laser.§*

3.2 Analyse des résultats

Tous types d'erreurs confondus, les catégories les plus spécifiques des contextes sources problématiques sont :

1. les participes passés (indice de spécificité : +4, fréquence totale dans l'original : F_i=322, fréquence locale dans les phrases comportant des erreurs dans la traduction en français : k_{ij}=277)
2. l'auxiliaire *be* au passé (+3, F_i=64, k_{ij}=59)
3. les modaux (+2, F_i=88, k_{ij}=77)
4. les prépositions (+2, F_i =1 620, k_{ij} =1 316)
5. les noms au pluriel (+2, F_i =915, k_{ij}=746)
6. la préposition *to* (+2, F_i=238, k_{ij}=198)
7. les gérondifs et participes présents (+2, F_i=175, k_{ij}=149).

Par exemple :

> *(Original) Fluid* ***inclusions*** *in the vein* ***minerals*** *representative of first breakdown* ***fluids*** *and fluid* ***inclusions*** *in olivine-enstatite* ***representing*** *final breakdown* ***fluids, were analysed*** *by* ***crushing*** *the* ***samples*** *in vacuum. §*
>
> *(Traduction) Les échantillons d'inclusions fluides dans les veines minérales représentatives de* ***la****[Type-annotateur_UD] première rupture des fluides, et de celles représentant* ***la_rupture****[Terme-traduit-par-non-terme] finale, ont été brisés* ***dans_la_chambre_pour_analyse****[Distorsion]. §*

De façon générale, les diagnostics de *spécificités* indiquent qu'il y a moins d'erreurs dans la traduction des énoncés comportant des séquences de nombres, unités de mesure, symboles, références, noms propres (-4). Les phrases avec les verbes (avec ou sans auxiliaire) au présent à la 3ème personne du singulier posent également moins de problèmes de traduction (-3), par exemple :

*(Original) Quartz **is** generally fine-grained and calcite **occurs** mostly as cement in the matrix (**Fig. 2a**).§*

(Traduction) Le quartz est généralement composé de grains fins et la calcite se trouve essentiellement comme ciment dans la matrice (Fig. 2a). §

Au total, les *spécificités* morpho-syntaxiques des contextes sources calculées au seuil fixé à 10 couvrent 2 699 occurrences de formes annotées. Les erreurs relevées dans l'annotation des productions totalisent 2 277 occurrences de formes annotées.

On note que les correspondances type *spécificités* sources/erreurs de traduction ne constituent pas des alignements parfaits au niveau sous-phrastique mais attirent l'attention sur des éléments caractéristiques de la langue source qui seraient à l'origine des problèmes de traduction. Par ailleurs, toutes les occurrences spécifiques (en gras ci-dessous) ne déclenchent pas systématiquement des erreurs ; on constate encore plusieurs types d'erreurs en co-occurrence qui partagent les mêmes contextes :

*(Original) In **addition**, the **most** intense **mass** peaks of the second **category** correspond mainly to dioxygenated molecules and **exhibit** a more distinctive **preference** of even **number** of **carbon** over odd **number** than those corresponding to other extended **recurring** molecular series.§*

(Traduction)
En_outre,_cette_seconde_catégorie_présente_des_pics_de_masses_dont_les_plus_intenses**[Type-annotateur_TR-UD]* correspondent à des molécules dioxygénées, montrant une préférence plus marquée pour **un nombre** [Omission_TR-OM]** pair **de**[Formulation-maladroite_LA-ST-AW] carbones que d'autres séries moléculaires récurrentes d'une large étendue.§*

(Commentaires de l'annotateur)
**Vérifiez si le sens de votre traduction est bien le même que dans "the most intense mass peaks of the second category correspond mainly"*
***Il n'est pas inutile de rester précis dans ce type de texte et de traduire aussi "over odd number"*
**** Il faut rendre la comparaison plus claire (cf. "than those")*

Pour analyser ce type de résultats dans une perspective contrastive, on mobilise la visualisation différentielle en contexte disponible dans *Le Trameur*.

3.3 Aides visuelles au repérage de la complexité en contexte

Dans *Le Trameur*, la *Lecture Textométrique Différentielle* (LTD) fournit des aides à la lecture contrastive de textes comparés appuyées par l'affichage synchrone des résultats de leur analyse textométrique parallèle (Patin et al., 2016). Les deux ensembles textuels sont affichés simultanément à l'écran pour faciliter les comparaisons. Les éléments caractéristiques sélectionnés par seuillage dans les contextes sources et cibles sont rendus « visibles » au fil des textes par un système de surlignage. Cette visualisation différentielle permet de cerner les facteurs qui sont à

l'origine de chaque type d'erreur de traduction, par l'examen contextuel des différentes causes et processus qui y sont liés.

La Figure 2 montre un extrait de la LTD générée par *Le Trameur* sur les phrases correspondant aux erreurs de syntaxe. Les résultats du calcul des *spécificités* morpho-syntaxiques dans les contextes sources couvrent 280 occurrences de formes graphique annotées. Les erreurs de syntaxe (70 occurrences de formes graphiques annotées) et les spécificités morpho-syntaxiques calculées sur les contextes originaux (280 occurrences de formes graphiques) sont surlignées en jaune. La visualisation attire l'attention sur la structure complexe des GN en anglais qui intègrent de nombreux termes composés (reflétée par la surreprésentation caractéristique des noms au singulier et au pluriel, cf. Table 1). La traduction en français nécessite dans ce cas la maîtrise des stratégies de traduction qui rétablissent les relations explicites entre les constituants dans la phrase au niveau micro-syntaxique (pour les apprenants, ces types de relations sont souvent peu explicites en anglais de spécialité). Ces diagnostics relevant de la complexité du texte scientifique en anglais peuvent déclencher des requêtes et vérifications dans les corpus comparables du domaine de spécialité, selon la méthodologie mise en place dans l'expérimentation (Kübler et al. 2016).

VOLET : EN	VOLET : FR
Thus, H 2 generation may be episodic depending on the rates of formation and destabilization of mineral surface layers during progressive waterrock interaction in an open system.§	Conclusion, la production de H 2 peut être réalisée en plusieurs stades, selon les taux de formation et de déstabilisation de couches superficielles minérales, pendant une interaction eau-roche progressive, au sein dun circuit ouvert. §
Therefore, the extraction of continental crust from this already-depleted reservoir (the EDR) cannot have greatly increased the Sm/Nd ratio of the MORB source; otherwise its 143Nd/144Nd value would have evolved to higher values than those observed for any terrestrial rock.§	Par conséquent, lextraction de croûte terrestre de ce réservoir, déjà appauvrit, ne peux avoir augmenté le rapport Sm/Nd de la source de basalte de dorsale médio-océanique, de façon conséquente. Sans quoi sa valeur en 143Nd/144Nd aurait évolué de manière plus importante que les valeurs observées sur nimporte quelle roche terrestre.§
the volume of mantle from which the continental crust was extracted must be large.§	le volume de croûte terrestre extrait du manteau se doit être inférieur à celui-ci.§
These mechanisms require a positive volume change of dehydration; otherwise, elevated pore pressures will not occur.§	Ces mécanismes requièrent un changement de volume positif de la déshydration; sans quoi il ne pourra pas y avoir de pressions interstitielles élevées.§
Above 6.5 GPa, the antigorite dehydration reaction had a negative Clapeyron slope and volume change of reaction.§	Au-dessus de 6,5 GPa, la réaction de la déshydratation de l'antigorite avait une pente de Clapeyron négative et un changement de volume de réaction négatif. §
Owing to the near-edge spectral characteristics (peak position, structure) for potential candidate (hydr)oxides (for example, ferrihydrite, haematite and goethite), we cannot uniquely identify the Fe(iii)bearing phase and henceforth use the term Fe(iii)-(hydr)oxides to indicate Fe bound to O and/or OH in a variety of crystal structures.§	En raison des caractéristiques spectrales près du seuil d'absorption (position maximale, structure) pour les candidats potentiels d'(hydr-) oxydes (tels que le ferrihydrite, l'hématite et la goethite), nous ne pouvons pas identifier définitivement la phase contenant du Fe(III) et utiliserons donc le terme (hydr-) oxydes de Fe(III) pour signaler l'existance d'un lien entre Fe et O et/ ou HO dans des diverses structures de cristaux. §
Hydrothermal organic reactions affect petroleum formation, degradation, and composition (2, 3), provide energy and carbon sources for deep microbial communities (4, 5), and may be important in the origin of life (6, 7).§	Les réactions organiques hydrothermales ont un effet sur la formation, la dégradation et la composition du pétrole, elles pourvoient de l'énergie et des sources en carbone pour des communautés microbiennes profondes et peuvent avoir leur importance dans l'origine de la vie. §
Basaltic lavas erupted at some oceanic intraplate hotspot volcanoes are thought to sample ancient subducted crustal materials.§	Des laves basaltiques en provenance de certains volcans à point chaud situés entre deux plaques océaniques constitueraient des échantillons d'anciens matériaux crustaux subduits. §
Here we report anomalous sulphur isotope signatures indicating mass-independent fractionation (MIF) in olivine-hosted sulphides from 20-million-year-old ocean island basalts from Mangaia, Cook Islands (Polynesia), which have been suggested to sample recycled oceanic crust.§	Dans cet article nous rapportons l'existence de signatures isotopiques atypiques du soufre présentes dans des sulfures hébergés dans des olivines en provenance de basaltes d'îles océaniques de Mangaia, Îles Cook (Polynésie), datant de 20 millions d'années. Ces signatures atypiques relèvent d'un phénomène de fractionnement indépendant de la masse (MIF) et constitueraient un échantillon de croute océanique recyclée.§
Terrestrial MIF sulphur isotope signatures (in which the amount of fractionation does not scale in proportion with the difference in the masses of the isotopes) were generated exclusively through atmospheric photochemical reactions until about 2.45 billion years ago.§	Les signatures isotopiques terrestres du sulfure produites par MIF (phénomène dans lequel le fractionnement n'est pas proportionnel à la différence de masse entre les isotopes) ont été générées exclusivement par réactions photochimiques atmosphériques qui ont eu lieu jusqu'il y a 2,45 milliards d'années. §
The RZ is composed of quartz, wollastonite (Wo, grey in Fig. 1a), Ca-rich garnet (Grt) and CM, and lacks calcite and phengite.§	La ZR est constituée de quartz, wollastonite (Wo, en gris dans la Fig. 1a), grenat roche en carbone (Grt), et matière carbonée. Elle ne contient pas de calcite, ni phengite. §
Using laser-heated diamond anvil cells, we constructed the solidus curve of a natural fertile peridotite between 36 and 140 gigapascals. §	À l'aide de cellules à enclume de diamants chauffées par laser, nous avons construit la courbe solidus d'une péridotite fertile et naturelle sous une pression allant de 36 à 140 Gigapascals. §

FIGURE 2: Spécificités sur contextes d'erreurs de syntaxe (export généré par *Le Trameur*)

4 Conclusion et perspectives

Nous avons avancé les premières propositions pour le profilage des erreurs de traduction en contextes alignés. La détection des difficultés d'apprentis traducteurs face aux textes originaux a mobilisé les annotations des productions selon la typologie d'erreurs MeLLANGE (Kübler et al., 2016) et la méthode des *spécificités* (Lafon, 1984 ; Lebart, Salem, 1994) avec la *Lecture Textométrique Différentielle* (LTD) sur contextes parallèles (Patin et al., 2016). Le repérage des éléments caractéristiques des contextes sources correspondant à un certain type d'erreur de traduction a permis de récolter des indices quantitatifs sur les constructions potentiellement complexes qui sont à l'origine des difficultés des apprenants (transfert de contenu, erreurs de langue, etc.).

Ces diagnostics peuvent être exploités de plusieurs façons. Ils peuvent alerter les apprenants eux-mêmes et donner lieu à des vérifications en corpus comparable du domaine de spécialité, mais aussi être utilisés par les enseignants pour proposer des exercices ciblés constitués à base de corpus. Cette voie ouvre des perspectives pour le développement des supports de cours informatisés qui allient la typologie d'erreurs issue de l'annotation des productions et l'exploration dynamique des contextes alignés et profilés par type d'erreurs.

Dans les expérimentations à venir, nous envisageons de comparer les erreurs de traductions dans les productions réalisées avec et sans apport de corpus (corpus *ER-TRAD-SP1* et *ER-TRAD-SP2*) afin d'observer si les *spécificités* des contextes sources liées aux erreurs changent en fonction des conditions de production des traductions. Ces nouveaux chantiers visent à alimenter la réflexion sur les méthodes d'enseignement qui amènent la diminution du nombre d'erreurs de traduction amorcée dans la première phrase d'expérimentation (Kübler et al., 2016).

Remerciements

Les auteurs remercient tous les membres de l'équipe CLILLAC-ARP (Paris 7), notamment Alexandra Mestivier (Volanschi) et Mojca Pecman, qui ont participé à l'annotation des erreurs de traduction et à la création des corpus utilisés dans ce volet de l'étude.

Références

BOWKER L., BENNISON P. (2003). Student Translation Archive and Student Translation Tracking System. Design, Development and Application. In Zanettin F., Bernardini S. and Stewart D. editors, *Corpora in translator education*. Manchester: St. Jerome Publishing.

CASTAGNOLI, S., CIOBANU D., KÜBLER N., KUNZ K., VOLANSCHI A. (2011). Designing a Learner Translator Corpus for Training Purposes. In Kübler N. editor, *Corpora, Language, Teaching, and Resources: From Theory to Practice*. Bern: Peter Lang.

Origines des erreurs en TS : différentiation textométrique grâce aux corpus de textes cibles annotés

FLEURY S., ZIMINA M. (2014). Trameur: A Framework for Annotated Text Corpora Exploration. Proc. of *COLING 2014 (the 25th International Conference on Computational Linguistics: System Demonstrations)*, August 2014, Dublin, Ireland, 57-61.

FRANKENBERG-GARCIA, A. (2015). Training translators to use corpora hands-on: challenges and reactions by a group of 13 students at a UK university. *Corpora*, 10/2, 351-380.

FRÉROT C. (2010). Outils d'aide à la traduction : pour une intégration des corpus et des outils d'analyse de corpus dans l'enseignement de la traduction et la formation des traducteurs. *Les Cahiers du GEPE* 2/2010, Outils de traduction - outils du traducteur ?

FROELIGER N. (2013). *Les Noces de l'analogique et du numérique - De la traduction pragmatique*. Paris : Les Belles lettres (collection Traductologiques).

GRANGER S., HUNG J., PETCH-TYSON S. (eds.) (2002). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam and Philadelphia: John Benjamins.

KÜBLER, N. (2011). Working with different corpora in translation teaching. In Frankenberg-Garcia A., Flowerdew L., and Aston G. editors, *New Trends in Corpora and Language Learning*. London: Continuum.

KÜBLER N, YVON F., WISNIEWSKI G. (2013). Human Errors and Automatic Errors in Machine Translations. What are the Differences? *Errare Workshop*, Ermenonville 2013.

KÜBLER N., MESTIVIER A., PECMAN M., ZIMINA M. (2016). Exploitation quantitative de corpus de traductions annotés selon la typologie d'erreurs pour améliorer les méthodes d'enseignement de la traduction spécialisée. Actes des *13èmes Journées internationales d'analyse statistique des données textuelles (JADT 2016)*, 7-10 juin, Nice, France.

LAFON P. (1984). *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion, Genève-Paris.

LEBART L., SALEM A. (1994). *Statistique textuelle*. Dunod.

LOOCK R., MARIAULE M., OSTER C. (2014). Traductologie de corpus et qualité : étude de cas, Actes du colloque *Tralogy II*, CNRS, 17-18 janvier, Paris, France.

PATIN S., ZIMINA M., FLEURY S. (2016). Lecture Textométrique Différentielle (LTD) de textes législatifs comparables de l'Union européenne. Actes des *13èmes Journées internationales d'analyse statistique des données textuelles (JADT 2016)*, 7-10 juin, Nice, France.

PEARSON J. (2003). Using parallel texts in the Translator Training Environment. In Zanettin F., Bernardini S. and Stewart D. editors, *Corpora in Translator Education*. Manchester: St Jerome Publishing.

WISNIEWSKI G., KÜBLER N., YVON F. (2014). A Corpora of Machine Translation Errors Extracted from Translation Students Exercises. Proc. of the *Ninth Language Resources and Evaluation Conference (LREC 2014)*, 26-31 May, Reykjavik, Iceland.

ZIMINA M. ET FLEURY S. (2015). Perspectives de l'architecture Trame/Cadre pour les alignements multilingues. *Nouvelles perspectives en sciences sociales : revue internationale de systémique complexe et d'études relationnelles* 11(1).

Patrons de coarticulation des voyelles françaises quantiques /i, a, u/ prononcées par des apprenants tchécophones. Illustration du logiciel VisuVo.

Nikola Maurová Paillereau
Laboratoire de Phonétique et Phonologie (UMR 7018), CNRS/ Paris 3, 19 rue des Bernardins, 75005 Paris, France
nikola.paillereau@mac.com

RESUME

Cette étude acoustique quantifie les écarts de prononciation des voyelles françaises quantiques /i, a, u/ produites par dix apprenants tchèques en isolation et dans des contextes symétriques labial pVp, dental tVt, palato-vélaire kVk et uvulaire RVR. Les productions de dix Françaises non-méridionales servent de référence. Les résultats, visualisés grâce aux schémas générés dans le logiciel VisuVo, montrent que 1. Les structures formantiques des voyelles produites par les apprenants diffèrent de la référence, notamment le F1 de [a] et le F2 de [i, a, u], et que 2. La réduction de l'aire des voyelles en contexte par rapport à celle des voyelles isolées est supérieure chez les apprenants comparées aux natives. L'ensemble des résultats indique que les voyelles quantiques ne sont pas réalisées par les deux groupes de locutrices avec des structures formantiques identiques, ce qui est en partie dû à des patrons de coarticulation différents.

ABSTRACT

Coarticulatory patterns of French quantic vowels /i, a, u/ as pronounced by Czech learners. Illustration of the VisuVo software.
This acoustic study quantifies pronunciation errors of French quantic vowels / i, a, u / produced by ten Czech learners in isolation and in symmetrical labial pVp, dental tVt, palato-velar kVk and uvular RVR contexts. The productions of ten non-southern French serve as reference. The results, displayed through diagrams generated in the software VisuVo show that 1. The formant structure of vowels produced by learners differs from the reference, especially the F1 of [a] and F2 of [i, a, u] and, 2. The reduction of the area of vowels in context relative to that of isolated vowels is higher in learners than in native speakers. The overall results indicate that quantic vowels are not realized by the two groups of speakers with identical formant, which is partly due to different patterns of coarticulation.

MOTS-CLES : tchèque, français langue étrangère (FLE), voyelles quantiques, formants, coarticulation, apprentissage.
KEYWORDS: Czech, French as a Foreign Language (FFL), quantic vowels, formants, coarticulation, learning.

1 Objets et évaluation de l'apprentissage phonétique

L'apprentissage phonétique d'une langue étrangère (L2) est un processus complexe. Il englobe les phénomènes suprasegmentaux, c'est-à-dire le rythme, l'accentuation lexicale et l'intonation, de même que les segments, c'est-à-dire les voyelles et les consonnes. Au plan segmental, l'apprenant apprend à maîtriser les phonèmes de la L2, leurs combinaisons possibles au sein de la syllabe définies par les règles phonotactiques de la L2 ainsi que les variations allophoniques (Best and Tyler 2007). Ces dernières dépendent entre autre de la position prosodique et du contexte phonétique. Les variations allophoniques suivant la position par rapport à la syllabe accentuée sont relativement petites dans des langues où les sons sont produits avec une forte tension musculaire, comme en français (Delattre 1969) ou bien où la durée, étant contrastive, rencontre de faibles variations dues à l'accentuation, comme en tchèque (Flemming 2005). En revanche, les variations dues au contexte phonétique caractérisent toute parole naturelle mais leur ampleur varie d'une langue à l'autre (Manuel et Krakow 1984).

Si l'apprentissage phonétique d'une L2 est jugé difficile, l'évaluation exacte du niveau phonétique en L2 n'en l'est pas moins (Dickerson 1975). De nos jours, l'évaluation est essentiellement subjective, car les enseignants se fient le plus souvent à leur capacité à percevoir le progrès des apprenants. Même si cette étape est importante car l'effet que la parole de l'apprenant aura sur l'auditeur est primordial dans la communication interhumaine, l'évaluation peut se faire également de manière objective à l'aide des analyses de paramètres acoustiques et articulatoires (Ashby 1990). L'avantage de l'analyse instrumentale réside en la possibilité de tracer toute évolution du système phonétique de l'apprenant, ainsi que de définir et visualiser les zones de difficultés qu'il faut travailler tout particulièrement.

1.1 Patrons de coarticulation

Dans la chaîne parlée, les sons ne sont prononcés de manière isolée que très rarement (en français, les sons isolés apparaissent par exemple dans l'énonciation des interjections *Euh…* ou *Oh !)* (Wioland 2005). La plupart du temps, les voyelles et consonnes se trouvent en contexte phonétique immédiat qui influence leur nature. En effet, du fait de l'inertie variable des organes articulatoire, les gestes articulatoires lors de la production d'une suite de sons se chevauchent et le résultat peut être visible dans le signal acoustique sous forme de transitions formantiques (Fant 1973). Ce phénomène s'appelle la *coarticulation* car les sons adjacents sont en effet articulés ensemble.

La coarticulation est en partie déterminée par les lois acoustiques universelles. Les mouvements (transitions) formantiques des voyelles coarticulées sont quantifiés par rapport à la cible, qui est le plus souvent définie par des voyelles produites en isolation (Miller 1981) ou en contexte glottal (Stevens and House 1963). Ainsi, dans une suite Consonne-Voyelle (CV), la direction des transitions formantiques dépend du lieu d'articulation de la voyelle et de la consonne. Lorsqu'il s'agit d'une voyelle antérieure comme par exemple le /e/ français, où le deuxième formant (F2) relativement élevé est essentiellement affilié à la courte cavité antérieure, ce dernier va baisser au contact de la consonne labiale /p/, allongeant la cavité antérieure, rester stable au contact de la consonne dentale /d/, ayant le locus dental vers 1800 Hz (Delattre 1969) ou en revanche augmenter au contact de la consonne palato-vélaire /k/, ayant un locus vers 3000 Hz pour les voyelles antérieures (Kewley-Port 1982). Si la coarticulation s'expliquait uniquement par les lois acoustiques, les patrons de coarticulations seraient identiques dans toutes les langues et ne

poseraient aucun problème pour un apprentissage phonétique d'une L2. Or ceci n'est pas le cas. En effet, les langues varient quant à l'ampleur de la coarticulation. Manuel et Krakow (1984) expliquent ces différences de coarticulation par les différences de la taille des systèmes phonologiques. Les auteurs proposent que les langues ayant des inventaires phonologiques réduits rencontrent de plus amples variations allophoniques que les langues avec des inventaires phonologiques plus riches. Même si cette hypothèse n'est pas toujours confirmée (Maddieson et Wright 1996), les études de coarticulation confirment que les langues du monde varient de part de l'ampleur de la coarticulation (pour revue de littérature, voir Steinlen 2005). Ces résultats, appliqués au niveau de l'apprentissage phonétique d'une L2, expliquent pourquoi les apprenants doivent entre autre acquérir les patrons de coarticulation spécifiques à la L2.

En 1995, le Speech Learning Model – SLM, qui est un modèle d'acquisition de la parole, a été clairement décrit par Flege. Le SLM a servi de cadre pour de nombreuses études acoustiques sur la production des segments d'une L2. Cependant, même si Flege (1995) attire l'attention sur la nécessité pour un apprenant de maîtriser toutes les variations allophoniques afin de créer une représentation cognitive d'un phonème, peu de chercheurs s'intéressent aux effets de coarticulation en L2. A notre connaissance, aucune étude concernant l'effet systématique des consonnes sur les voyelles du FLE n'a été réalisée jusqu'à présent.

1.2 Les systèmes vocaliques du français et du tchèque

Le français, langue romane, est caractérisé par un système vocalique riche. Selon le pays et la région où le français est parlé comme langue maternelle (L1), le système vocalique est composé de trois à quatre voyelles nasales ainsi que de sept à onze voyelles orales (Vaissière 2006). Ainsi, le système maximal comporte les voyelles /i, e, ɛ, a, y, ø, œ, u, o, ɔ, ɑ, ɶ̃, ɛ̃, ɔ̃, ɑ̃/ alors que le système minimal, caractéristique du français méridional, élimine les oppositions entre les voyelles mi-fermées/ mi-ouvertes e/ɛ, o/ɔ, ø/œ, de même que le /ɑ/ postérieur, pour opposer /ɑn/ « âne » de /an/ « Anne » ou le /ɶ̃/ pour opposer /bRɶ̃/ « brun » de /bRɛ̃/ « brin » (Léon et Léon 2007). Le français standard - important à définir pour les besoins de la phonétique didactique - est parfois décrit comme le français « non méridional » (Pustka 2011). Le système vocalique du français standard contient dix voyelles orales et trois voyelles nasales qui doivent donc être présentées aux apprenants de français langue étrangère (FLE). Cependant, il existe une hiérarchie entre les voyelles du français qu'il faut prendre en compte dans l'apprentissage phonétique (Wioland 2005). Du fait de la distribution semi-complémentaire des voyelles moyennes et du nombre restreint de paires minimales basées sur leur opposition, les contrastes e/ɛ, o/ɔ et ø/œ ont un poids fonctionnel plus petit que les contrastes avec les voyelles fermées /i, y, u/ et la voyelle ouverte /a/.

Le tchèque, langue slave, est caractérisé par un système vocalique plus restreint. Traditionnellement, l'on postule cinq qualités vocaliques avec un contraste de durée, opposant les voyelles brèves /a, e, i, o, u/ et longues /aː, eː, iː, oː, uː/ (Kučera et George 1968). Les études instrumentales plus récentes pointent sur l'ouverture des voyelles fermées brèves (Skarnitzl et Volín 2012). Les auteurs tchèques contemporains décrivent ainsi le système phonologique des locuteurs du tchèque de Bohème (qui correspond au tchèque standard) avec le contraste entre les voyelles longues tendues et les voyelles brèves relâchée iː/ɪ (Šimáčková et al. 2012) et parfois également le contraste entre uː/ʊ (Duběda 2005). De la même façon, suite aux études instrumentales, les symboles pour indiquer une qualité mi-ouverte sont de plus en plus utilisés pour la transcription des voyelles moyennes antérieures /ɛ, ɛː/ et parfois pour celle des voyelles moyennes postérieures du

tchèque /ɔ, ɔː/ (Duběda 2005). Enfin, le système tchèque comporte également trois voyelles diphtongues /au, eu, ou/.

Nous avons récemment réalisé (Paillereau, à paraîtreB) une comparaison des formants vocaliques des voyelles orales du français et des voyelles monophtongues du tchèque produites isolément et en contextes symétriques labial, dental et palato-vélaire. Les résultats montrent que les deux langues varient par l'ampleur de la coarticulation, comme attendu. Contre attente cependant, les voyelles postérieures du français sont plus fortement coarticulées, notamment en contexte dental, que les voyelles tchèques, moins nombreuses. Ces résultats peuvent être appliqués en didactique de la prononciation du FLE pour les apprenants tchécophones où des difficultés au niveau de la coarticulation des voyelles peuvent être anticipées du fait du transfert de la L1 vers la L2.

1.3 VisuVo, logiciel de visualisation des formants vocaliques

Dans notre thèse de doctorat (Paillereau 2015), nous avons élaboré le logiciel VisuVo pour diagnostiquer les difficultés que les apprenants rencontrent dans la production des voyelles d'une L2. Le logiciel est composé de deux parties : la partie Serveur qui permet d'analyser en moins de deux secondes les données brutes, en calculant la moyenne et l'écart type des formants. Ensuite, la partie Client, écrite en Javascript, qui génère instantanément des schémas de visualisation. Les graphiques présentés dans cet article permettent de rendre compte des patrons de coarticulation en traçant l'évolution des formants vocaliques suivant la position dans le mot (initiale, médiane et finale) et le contexte consonantique symétrique (consonnes de différents lieux d'articulation : labiale /p/, dentale /t/, palato-vélaire /k/ et uvulaire /R/).

L'avantage du logiciel VisuVo par rapport aux logiciels comme Praat (Boersma et Weenink 2015) réside en la rapidité de la visualisation des résultats et en son interactivité. Il est possible de comparer différentes variables, comme les voyelles françaises et tchèques telles qu'elles sont produites par des natifs, ou bien des voyelles françaises produites par des natifs et par des apprenants, ou bien observer les différences individuelles en comparant les voyelles produites par les différents apprenants. Selon les objectifs, il revient ainsi à l'utilisateur d'organiser ses données dans un fichier Excel, qui sert d'entrée au logiciel VisuVo.

2 Objectifs

Le but de l'article est d'examiner les effets systématiques des consonnes de différents lieux d'articulation – labial /p/, dental /t/, palato-vélaire /k/ et uvulaire /R/ - sur les voyelles quantiques du français /i, a, u/ produites par des apprenants tchécophones et de comparer les données à la référence. Les voyelles quantiques /i, a, u/ sont de l'intérêt particulier car elles sont définies par Stevens (1989) comme acoustiquement stables. Cependant, de nombreuses études ont montré que les structures formantiques des voyelles /i, a, u/ varient d'une langue à l'autre (Steinlen 2005, Gendrot et Adda-Decker 2007).

Les apprenants tchécophones du niveau avancé produiront-ils les voyelles françaises /i, a, u/ avec des structures formantiques différents de ceux de la « référence » ? Les écarts en production dépendront-ils des contextes consonantiques dans lesquelles les voyelles apparaissent ? La réalisation des patrons de coarticulation sera élucidée à l'aide des graphiques générés dans VisuVo.

3 Méthodologie

3.1 Corpus

Le corpus est celui de Landron et al. (2010). Les voyelles analysées dans cette étude sont le /a/, /i/ et /u/ du français. Elles apparaissent :

– En isolation, séparées par des petites pauses, dans des phrases cadre du type : « Bébé, il a dit <é> comme dans bébé ».
– En contexte consonantique symétrique, dans des logatomes trisyllabiques CVCVCVC où C correspond aux consonnes de quatre lieux d'articulation différents : labial /p/, dental /t/, palato-vélaire /k/ et uvulaire /R/. Les logatomes sont insérés dans des phrases cadre telles que « Le mot *kaukaukauke* peut bien coller. »

Toutes les phrases ont été présentées aux locuteurs quatre fois dans un ordre prédéfini de manière à ce qu'une même phrase ne soit jamais répétée à la suite. Avant les enregistrements, les locuteurs ont d'abord suivi une phase d'entraînement avec des consignes écrites et des exemples sonores préenregistrés.

3.2 Locuteurs

Les phrases ont été lues par dix Françaises natives dont les productions ont servi à définir la « référence » acoustique. Des femmes plutôt que des hommes ont été choisies pour des rasions pratiques : la plupart des étudiants du français à l'université, comparées à la « référence », sont des femmes. Les locutrices natives sont âgées entre 21 ans et 48 ans (M âge = 28,5 ; E.T. = 7,5). Aucune locutrice n'est née ni n'a grandi dans le sud de la France.

Le groupe test est composé de dix locutrices tchèques, âgées entre 25 ans et 28 ans (M âge = 25,7 ; E.T. = 1,3). Toutes les locutrices sont nées et ont grandi dans la région de Bohème et elles ont commencé à apprendre le français après l'âge de onze ans. Au moment des enregistrements, les locutrices étaient étudiantes en Master de FLE à l'université de Bohème et possédaient un bon niveau global du français (B2/C1).

3.3 Conditions d'enregistrement et traitement des données

Les enregistrements ont été réalisés dans une pièce calme (chambre sourde pour les locutrices natives et une pièce silencieuse pour les apprenants tchécophones) avec une carte son Edirol UA 25, reliée à un ordinateur Apple et un microphone serre-tête AKGC 520 L. L'utilisation du microphone serre-tête permet d'assurer une distance constante entre les lèvres et le microphone tout le long de l'enregistrement. La fréquence d'enregistrement était fixée à 44 100 Hz avec une résolution de 16 bits. Nous avons utilisé Audacity (http://audacity.sourceforge.net) pour enregistrer les phrases lues et Praat pour l'analyse acoustique.

Seulement les parties du signal acoustique comportant les voyelles isolées et les voyelles insérées dans les logatomes trisyllabiques ont été soumises à l'analyse. Les fichiers son ont été segmentés et étiquetés automatiquement et vérifiées manuellement. Le début et la fin vocalique coïncident avec l'apparition des formants et notamment du F2. Les valeurs des quatre premiers formants ont été relevées de manière semi-automatique avec la formule To Formant (burg) basée sur une analyse LPC[1], à environ un tiers, à la moitié et à deux tiers de la durée vocalique. La fenêtre d'analyse spectrale étant fixée à 25 ms (spectre à bandes larges) permet une bonne analyse temporelle. En cas de mauvaise détection des formants - souvent due à la voix craquée[2] ou au *voice decay time*[3] - retournant des valeurs aberrantes, les données ont été vérifiées et corrigées manuellement. Toutes les données brutes ont été stockées dans un fichier Excel qui a servi d'entrée au logiciel VisuVo pour la visualisation des données.

La voyelle postérieure /u/, caractérisée par une proéminence perceptive des basses fréquences (Delattre 1951), est définie par la valeur des deux premiers formants. La voyelle antérieure /a/ est définie par la valeur des trois premiers formants (comme les autres voyelles antérieures du français dont F1 à F3 sont visualisés dans VisuVo afin de rendre compte du contraste entre les voyelles étirées et arrondies, même si ce contraste n'est pas pertinent pour le /a/). Enfin la voyelle /i/ est définie par la valeur des quatre premiers formants. Dans la production du /i/ français, l'énergie est concentrée au niveau des troisième et quatrième formants, qui sont rapprochés et par conséquent amplifiés (Vaissière 2009). Cette caractéristique acoustique du /i/ n'est cependant pas partagée par toutes les langues, comme démontré par Gendrot et Adda-Decker (2007) qui ont comparé des inventaires vocaliques de huit langues différentes, y compris le français.

4 Résultats

4.1 Comparaison de la structure formantique de [i, a, u]

Afin de comparer la structure formantique selon le contexte et le groupe de locutrices, des analyses de variance ANOVA à trois facteurs ont été effectuées sur les mesures de F1 et F2 de toutes les voyelles, sur le F3 de /i/ et de /a/ de même que sur le F4 de /i/. Le facteur Contexte possède cinq niveaux : isolé (iso), labial (/p/), dental (/t/), palato-vélaire (/k/) et uvulaire (/R/) et le facteur Locuteurs possède deux niveaux : locutrices françaises natives (FR) et apprenantes tchécophones (TCH). Le facteur Voyelle possède trois niveaux dans l'analyse des formants F1 et F2 : /i/, /a/ et /u/ ; deux niveaux pour l'analyse du F3 : /i/ et /a/ ; et enfin le F4 n'est analysé que dans le cas de /i/.

Les Tableaux 1 et 2 indiquent les valeurs moyennes (en Hz) des formants et les écart types (en italique) pour chaque voyelle et chaque contexte séparément. Le Tableau 1 indique les données des dix locutrices françaises natives et le Tableau 2 les données des dix apprenants tchécophones. Les valeurs ayant servi pour le calcul de la moyenne sont les trois valeurs prises pour les voyelles isolées et la valeur prise à la moitié de la durée vocalique pour les voyelles coarticulées.

[1] Linear Predictive Coding est une estimation du signal basée sur des coefficients de prédiction qui caractérisent des formants.

[2] Caractérisée par une irrégularité de la période et de l'amplitude des vibrations des plis vocaux.

[3] Période de temps où les plis vocaux continuent à vibrer mais le son vocalique n'est plus audible.

Le résultat de l'analyse de variance montre la significativité de l'effet global ainsi que de l'effet de chacun des trois facteurs sur les formant F1 à F3 ($p<0,05$). En revanche, l'effet du facteur Locuteur sur la valeur de F4 n'est pas significatif ($p=0.47$). Le test à posteriori de Fisher indique à quels groupes est dû l'effet de l'interaction des trois facteurs au niveau de chaque formant. Pour le F1, la significativité est due à la différence ($p<0,05$) dans la production de /a/ (dans tous les contextes sauf palato-vélaire), de /i/ (en contexte labial et uvulaire) et de /u/ (en contexte uvulaire).

FR	Iso			pVp			tVt			kVk			RVR		
	i	a	u	i	a	u	i	a	u	i	a	u	i	a	u
F1	272	805	**287**	**275**	843	**299**	**296**	817	**298**	**277**	774	**295**	431	852	333
	26	*132*	*30*	*28*	*93*	*31*	*29*	*113*	*27*	*23*	*113*	*26*	*46*	*104*	*47*
F2	2524	1300	770	2360	1583	845	2417	1770	1176	2517	1873	**825**	**2499**	1412	**726**
	272	*178*	*72*	*266*	*110*	*72*	*266*	*85*	*174*	*215*	*117*	*81*	*228*	*128*	*81*
F3	3787	**2834**		**3513**	**2797**		**3455**	**2856**		3686	**2642**		**3318**	2785	
	229	*168*		*256*	*162*		*242*	*176*		*167*	*185*		*209*	*227*	
F4	**4423**			**4367**			**4334**			**4359**			**4238**		
	196			*208*			*195*			*161*			*311*		

TABLEAU 1 : Valeurs formantiques moyennes (en Hz) des voyelles françaises /i, a, u/ produites en isolation (iso) et en syllabe finale pVp, tVt, kVk et RVR par 10 Françaises natives (valeurs relevées à la moitié de la durée vocalique, 4 répétitions). L'écart type (en Hz) est en italique. Les formants identiques aux voyelles des apprenants sont en gras.

TCH	Iso			pVp			tVt			kVk			RVR		
	i	a	u	i	a	u	i	a	u	i	a	u	i	a	u
F1	253	839	**286**	**264**	806	**303**	**270**	764	**300**	**268**	775	**306**	379	799	364
	29	*87*	*33*	*30*	*67*	*37*	*22*	*78*	*43*	*24*	*70*	*39*	*65*	*90*	*44*
F2	2663	1479	708	2616	1476	727	2621	1635	1074	2697	1665	**777**	**2516**	1564	**787**
	249	*222*	*93*	*226*	*166*	*73*	*196*	*181*	*201*	*220*	*180*	*91*	*130*	*163*	*142*
F3	3711	**2800**		**3553**	**2766**		**3441**	**2913**		3597	**2594**		**3254**	2810	
	237	*119*		*153*	*163*		*166*	*160*		*213*	*207*		*253*	*153*	
F4	**4359**			**4364**			**4383**			**4379**			**4264**		
	209			*269*			*261*			*232*			*231*		

TABLEAU 2 : Valeurs formantiques moyennes des voyelles françaises /i, a, u/ produites en isolation (iso) et en syllabe finale pVp, tVt, kVk et RVR par 10 apprenants tchécophones (valeurs relevées à la moitié de la durée vocalique, 4 répétitions). L'écart type (en Hz) est en italique. Les formants identiques aux voyelles des Françaises natives sont en gras.

Au niveau du F2, la significativité est due aux voyelles produites dans la plupart des contextes. En effet, seules les voyelles /u/ (en contexte palato-vélaire et uvulaire) et /i/ (en contexte uvulaire) sont produites avec la même valeur moyenne du F2 par les deux groupes de locutrices. La différence significative de F3 des voyelles /i/ et /a/ selon le contexte et les locutrices est due à la voyelle /i/ produite en isolation et en contexte palato-vélaire, ainsi qu'à la voyelle /a/ produite en contexte uvulaire. Enfin, la voyelle /i/ a été produite avec la même valeur moyenne de F4 par les deux groupes de locutrices dans tous les contextes étudiés. Ces résultats sont en partie reportés dans les Tableaux 1 et 2 où les formants moyens identiques ($p>0,05$) pour une même voyelle produite par les deux groupes de locutrices sont en gras. Les Figures 3 à 5 montrent les patrons de coarticulation des voyelles françaises /i, a, u/ produites en isolation (traits horizontaux traversant l'image) et dans des logatomes CVCVCVC où C correspond respectivement à la consonne labiale, dentale, palato-vélaire et uvulaire. Les moyennes affichées des voyelles coarticulées sont celles des valeurs prises à un tiers, à la moitié et à deux tiers de la durée vocalique. Dans notre étude, nous nous intéressons aux moyennes des valeurs prises à la moitié de la durée des voyelles coarticulées finales. Pour les voyelles isolées, une valeur moyenne globale est calculée à partir des trois valeurs. Les voyelles des

Françaises natives sont marquées avec un rond et celles des apprenants tchèques avec une croix. L'écart-type affiché est de un.

4.1.1 Structure formantique du [i]

La Figure 3 illustre l'évolution des formants moyens F1 à F4 du [i] selon le contexte et les locutrices. En isolation, les apprenants produisent le [i] par rapport à la référence avec un F1 plus bas (253 Hz contre 272 Hz), un F2 plus élevé (2663 Hz contre 2524 Hz) et un F3 plus bas (3711 Hz contre 3787 Hz) (pour l'ensemble des résultats, se référer aux Tableaux 1 et 2). Lorsque [i] apparaît en contexte labial, dental et palato-vélaire, les Tchécophones le produisent avec un F2 plus élevé (de 256 Hz, 204 Hz et 180 Hz respectivement) que la référence. En contexte uvulaire, le [i] des apprenants est réalisé avec un F1 plus bas que la référence (379 Hz contre 431 Hz). Cette différence est probablement due à une différence dans l'articulation de la consonne /R/ que les Tchèques réalisent parfois comme apicale (Nováková, à paraître).

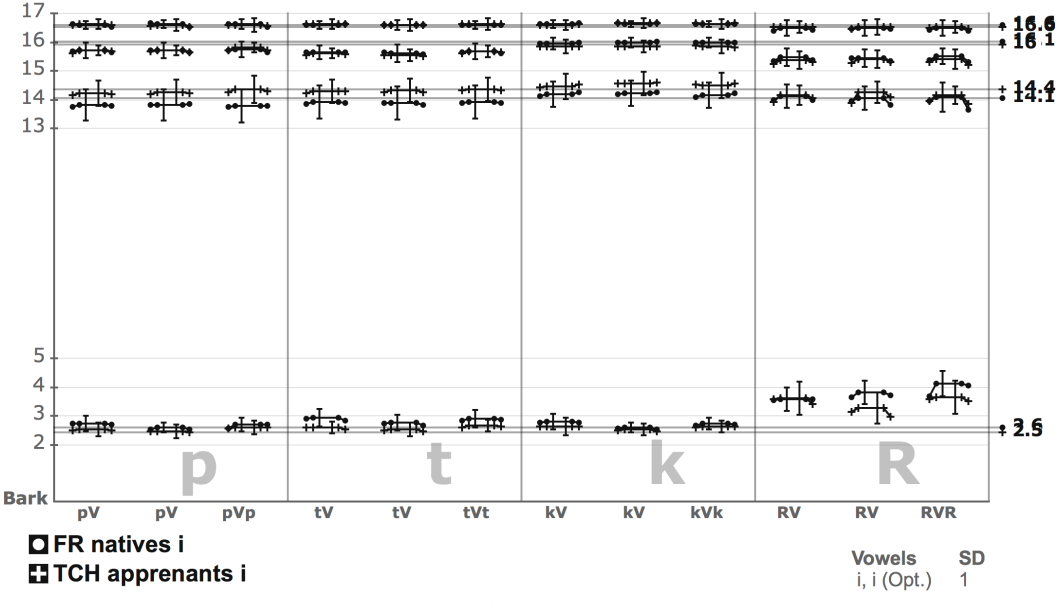


FIGURE 3: Formants moyens F1, F2, F3 et F4 (en Bark) du [i] français produit par 10 Tchèques (croix) et par 10 Françaises (rond) en isolation (traits horizontaux) et dans des logatomes CVCVCVC où C = [p, t, k, ʁ]. L'écart-type est de un.

4.1.2 Structure formantique du [a]

La Figure 4 illustre l'évolution des formants moyens F1 à F3 du [a] selon le contexte et les locutrices. En isolation, les apprenants produisent le [a] par rapport à la référence avec un F1 et un F2 plus élevés (respectivement 839 Hz contre 805 Hz et 1479 Hz contre 1300 Hz). En contexte labial et dental, la tendance s'inverse car les apprenants produisent le [a] avec un F1 et F2 plus bas que la référence. Les valeurs sont respectivement de 806 Hz et 1476 Hz contre 843 Hz et 1583 Hz en contexte labial et de 764 Hz et 1635 Hz contre 817 Hz et 1770 Hz en contexte dental. En contexte palato-vélaire, la différence concerne le deuxième formant uniquement : les apprenants produisent un [a] avec un F2 de 208 Hz plus bas que la référence. Ce résultat soutient une forte

antériorisation acoustique du [a] français produit en contexte consonantique (Paillereau, à paraîtreA) qui est ici une source de problèmes pour les apprenants. Enfin, en contexte uvulaire, le [a] des apprenants est réalisé avec un F1 plus bas (799 Hz contre 852 Hz) et un F2 plus élevé (1564 Hz contre 1412 Hz) que la référence, ce qui peut de nouveau s'expliquer par l'articulation apicale de certaines occurrences du /R/ produit par les Tchèques.

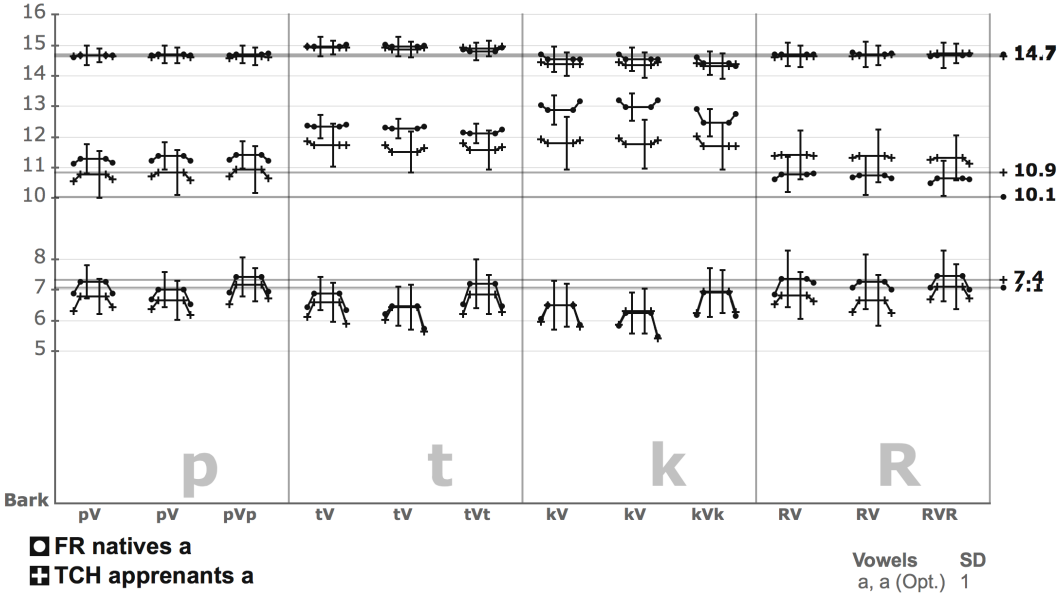


FIGURE 4: Formants moyens F1, F2 et F3 (en Bark) du [a] français produit par 10 Tchèques (croix) et par 10 Françaises (rond) en isolation (traits horizontaux) et dans des logatomes CVCVCVC où C = [p, t, k, ʁ]. L'écart type est de un.

4.1.3 Structure formantique du [u]

L'évolution des formants F1 et F2 du [u] français selon le contexte et les locutrices est illustrée à la Figure 5. En isolation, les apprenants produisent le [u] français avec un F2 plus bas que la référence (708 Hz contre 770 Hz). Ce résultat est surprenant compte tenu du fort caractère focal caractérisant le [u] français, qui est un parfait exemple des voyelles cardiales telles que définies par Daniel Jones (Vaissière 2009). Cette différence va dans le même sens en contexte labial et dental dans lequel les Tchèques produisent le [u] avec un F2 plus bas que les Françaises (il est de 727 Hz contre 845 Hz pour le contexte labial et de 1074 Hz contre 1176 Hz pour le contexte dental). En contexte palato-vélaire, le [u] des apprenants est réalisé avec exactement la même structure formantique que le [u] des Françaises natives ($p>0,05$ pour F1 et F2). Enfin, en contexte uvulaire, la différence concerne le F1 qui est de 31 Hz plus élevé chez les apprenants que chez les natives.

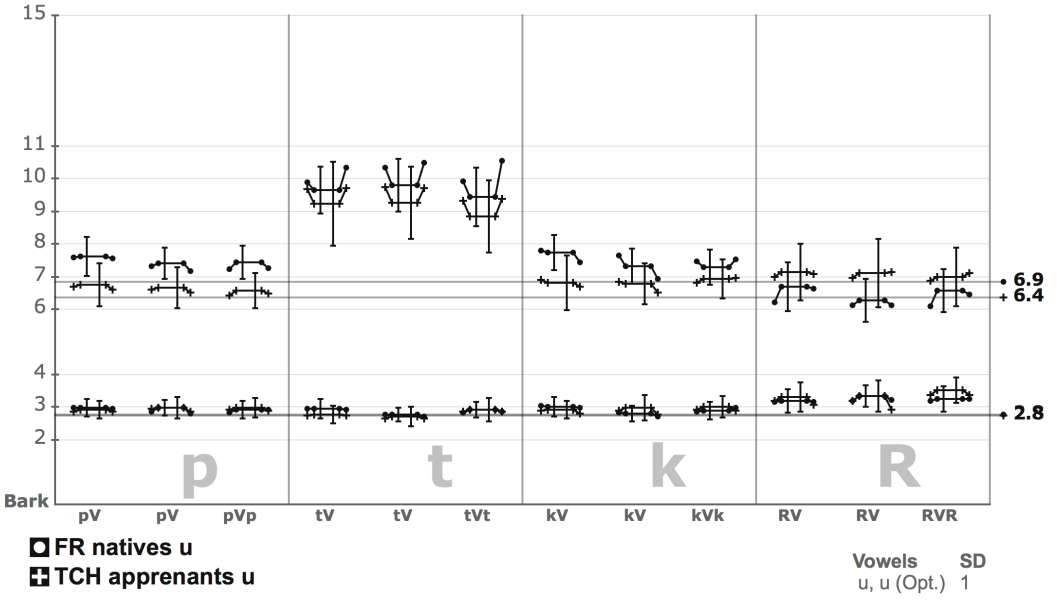


FIGURE 5: Formants moyens F1 et F2 (en Bark) du [u] français produit par 10 Tchèques (croix) et par 10 Françaises (rond) en isolation (traits horizontaux) et dans des logatomes CVCVCVC où C = [p, t, k, ʁ]. L'écart-type est de un.

4.2 Aire du triangle vocalique /i, a, u/

Afin d'examiner l'effet du contexte sur la réduction vocalique, l'aire de l'espace vocalique des locutrices françaises natives et des apprenants tchèques est également comparée. La taille de cette aire, exprimée en Hz^2, est calculée selon la formule de Heron. Dans un premier temps est calculée la moitié du périmètre du triangle *s* :

$s = (a+b+c)/2$,

puis est calculée l'aire *A* du triangle :

$A = \sqrt{s\,(s-a)\,(s-b)\,(s-c)}$.

Le Tableau 6 indique la taille de l'aire (en Hz^2) en fonction du contexte dans lequel les voyelles sont prononcées (en isolation et en syllabe finale pVp, tVt, kVk et RVR). Puisque la valeur absolue de la taille de l'aire n'a pas de sens en soi mais doit être interprétée en fonction de la taille de l'aire des voyelles isolées qui servent de cible (Steinlen 2005), le Tableau 6 indique également la différence de taille de l'aire (en Hz^2 et en %) des voyelles en contexte et de celle des voyelles isolées.

	FR			TCH		
	Aire (Hz^2)	Diff. Hz^2 Iso-CVC	Diff. % Iso-CVC	Aire (Hz^2)	Diff. Hz^2 Iso-CVC	Diff. % Iso-CVC
Iso	458.261			553.279		
pVp	420.936	- 37.325	- 8%	489.689	-63.590	-11%
tVt	322.634	- 135.627	- 30%	367.319	- 185.960	- 34%
kVk	414.666	- 43.595	- 9%	467.112	- 86.167	- 16%
RVR	426.480	- 31.781	- 7%	370.230	- 183.049	- 33%

TABLEAU 6 : La taille des aires (en Hz^2) et la différence des aires des triangles acoustiques (en Hz^2 et en %) calculées à partir des voyelles /i, a, u/ produites par 10 Françaises natives (FR) et 10 apprenants tchécophones (TCH) en isolation (iso) et en syllabes finales pVp, tVt, kVk et RVR.

Le Tableau 6 met en évidence une réduction de l'aire des voyelles produites en contexte par rapport à l'aire des voyelles isolées, comme attendu. Le degré de réduction n'est cependant pas identique pour les deux groupes de locuteurs. En contexte labial, l'aire vocalique des apprenants est réduite de 11% alors que celle des natives est réduite de 8%. En contexte dental et palato-vélaire, elle est réduite de 34% et 16% respectivement chez les apprenants et de 30% et 9% respectivement chez les natives. Enfin, en contexte uvulaire l'aire entre les voyelles /i, a, u/ est réduite de 33% dans la production des apprenants et de 7% seulement dans la production des Françaises natives. La différence de réduction d'aire en contexte uvulaire est illustrée à la Figure 7 qui est un triangle F1/F2 des voyelles /i, a, u/ produites en isolation (trait plein) et en syllabe finale RVR (trait pointillé) par des Françaises natives (à gauche) et des tchécophones (à droite). Ainsi, dans tous les contextes consonantiques étudiés, l'aire vocalique est davantage réduite dans la production de /i, a, u/ par des apprenants que par des natives.

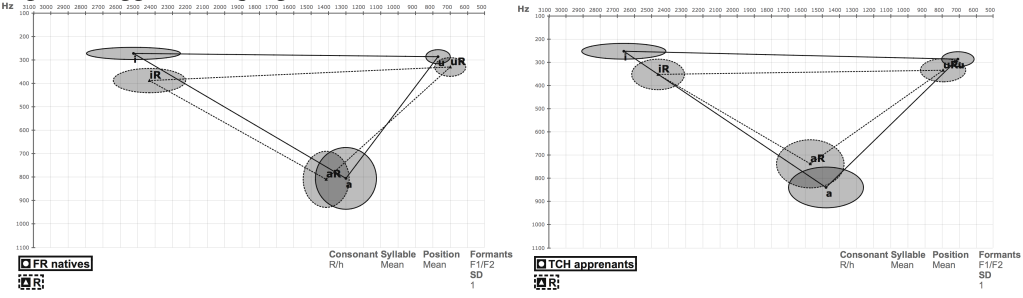


FIGURE 7: L'aire du triangle acoustique (F1/F2) des voyelles /i, a, u/ produites en isolation (trait plein) et en syllabe finale RVR (trait pointillé) par 10 Françaises natives (FR, à gauche) et par 10 apprenants tchécophones (TCH, à droite). L'écart type est de un.

5 Conclusion

L'étude acoustique des voyelles quantiques françaises /i, a, u/ a mis en évidence que les apprenants tchécophones ne produisent pas ces dernières avec les mêmes structures formantiques que les Françaises natives, mis à part le [u] en contexte palato-vélaire. Ce résultat est contradictoire à la stabilité acoustique de ces voyelles clamée par la théorie quantique de Stevens (1989) mais soutient de nombreuses études inter-langues (par exemple Gendrot et Adda-Decker 2007). Les écarts en production concernent essentiellement le F1 de [a], en partie corrélé à l'aperture du conduit vocal, et le F2 de toutes les voyelles, en partie corrélé au lieu d'articulation antérieur-postérieur de la voyelle. Ainsi, les apprenants du français doivent viser un F1 plus élevé du [a] en contexte, qui peut être atteint par une plus grande ouverture de la cavité buccale, et un F2 plus élevé des voyelles /u/ et /a/ en contexte dental, qui peut être atteint par une articulation plus antérieure. Dans la production du [a] isolé, les apprenants doivent en revanche parvenir à rapprocher le F1, qui est élevé, et le F2, qui est bas. Le caractère focal du [a̲] isolé caractérise en effet le français (Vaissière 1985). Dans la production du [i], les apprenants doivent viser un troisième formant maximal qui peut être atteint par une articulation très antérieure. Il est important de rappeler l'existence de compensations articulatoires qui font qu'une même structure formantique peut être obtenue avec des configurations articulatoires différentes (Maeda, 1989). Les stratégies articulatoires présentées ci-dessus peuvent

donc être proposées aux apprenants mais ne doivent pas être considérées comme les seules configurations possibles.

Nous avons également mis en évidence que les zones de difficultés dépendent des contextes consonantiques. Dans certains contextes, les écarts des apprenants par rapport à la référence sont plus grands que dans d'autres contextes. En prenant l'exemple du /u/ qui est le plus parlant, les résultats montrent que les apprenants produisent cette voyelle de manière native en contexte palato-vélaire alors qu'ils devraient viser une valeur du F2 plus élevée de la voyelle produite en isolation et en contextes labial et dental.

Les résultats concernant la production des voyelles en contexte uvulaire suggèrent que les apprenants produisent certaines occurrences du /R/ avec un lieu d'articulation apical, comme pour le /r/ tchèque, et non pas uvulaire, qui est le lieu d'articulation standard de ce phonème en français (Galazzi, à paraître). Malgré l'absence de l'étude articulatoire, ce phénomène est indiqué par un abaissement du F1 et une augmentation du F2 des voyelles /u/ et /a/ en contexte /R/ par rapport à leur cible. Si l'articulation était uvulaire, comme chez les Françaises natives de cette étude, le F1 devrait augmenter et F2 devrait baisser selon les lois acoustiques (Stevens, 1998).

La comparaison des aires vocaliques montre qu'une plus grande réduction de l'espace concerne les voyelles /i, a, u/ produites en contexte dental. Cependant, une réduction d'un tiers environ est également observée pour les voyelles produites en contexte /R/ par les apprenants, mais non pas par les Françaises natives, ce qui peut de nouveau s'expliquer par une réalisation différente de la consonne.

Enfin, il faut bien garder à l'esprit que même si les données acoustiques apportent des résultats concernant des zones de difficulté des apprenants, une vérification perceptive des voyelles produites devrait également être conduite. En effet, quoique les tests statistiques montrent une différence significative entre les formants des voyelles produites par des apprenants et des locutrices natives, il est possible que ces différences ne soient pas toujours perçues. Les graphes de visualisation des résultats montrent que les voyelles /i, a, u/ sont dans la plupart des contextes réalisées par des apprenants avec des formants à un écart type de la référence, ce qui correspondrait selon des critères appliqués dans certaines études acoustiques aux productions à priori fidèles (Birdsong, 2003).

Références

ASHBY, M.G. (1990). Prototype categories in phonetics. *Speech Hearing and Language :Work in Progress U.C.L.* 4, 21-28.

GALAZZI, E. (à paraître). Du locuteur natif à l'étranger expert: quel(s) modèle(s) de prononciation pour les apprenants de FLE dans la société globalisée? In M. Bořek-Dohalská, K. Suková Vychopňová (Eds.), *Didactique de la phonétique et phonétique en didactique du FLE*. Praha : Karolinum.

BEST, C. T., TYLER, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O. S. Bohn (Eds.), *Second language speech learning: The role of language experience in speech perception and production,* 13-34. Amsterdam: John Benjamins.

BIRDSONG, D. (2003). Authenticité de prononciation en français L2 chez des apprenants tardifs anglophones : analyses segmentales et globales. *Acquisition et interaction en langue étrangère,* 17-36.

BOERSMA, P., & WEENINK, D. (Producer). (2015). Praat: doing phonetics by computer.

DELATTRE, P. (1951). The Physiological Interpretation of Sound Spectograms. *PMLA* 66(5), 864-875.

DELATTRE, P. (1969). An Acoustic and Articulatory Study of Vowel Reduction in Four Languages. *International Review of Applied Linguistics in Language Teaching* 7(4), 295-324.

DICKERSON, L. B. (1975). The learner's interlanguage as a system of variable rules. *TESOL Quarterly* 9, 401-407.

DUBĚDA, T. (2005). *Jazyky a jejich zvuky : univerzálie a typologie ve fonetice a fonologii*. Praha: Karolinum.

FANT, G. (1973). *Speech sounds and features*. Cambridge: MIT Press.

FLEGE, J. E. (1995). Second language speech learning. Theory, findings, and problems. In W. Strange (ed) *Speech Perception and Linguistic Experience : Theoretical and Methodological Issues in Cross-Language Speech Research*, 233-272. Timonium, MD: York Press Inc.

FLEMMING, E. (2005). *A phonetically-based model of phonological vowel reduction*. ms. MIT.

GENDROT, C., ADDA-DECKER, M. (2007). Impact of duration and vowel inventory size on formant values of oral vowels: an automated formant analysis from eight languages. Actes de *Inter. Con. Phon. Sciences*, Saarbrûcken: 1417-1420.

KUČERA, H., GEORGE, K. M. (1968). *A comparative quantitative phonology of Russian, Czech and German*. New York: American Elsevier Publishing Company.

KEWLEY-PORT, D. (1982). Measurement of formant transitions in naturally produced stop consonant vowel syllables. *Journal of the Acoustic Society of America* 72, 379-389.

LANDRON, S., PAILLEREAU, N., NAWAFLEH, A., EXARE, C., ANDO, H., GAO, J. (2010). Le corpus PhoDiFLE : un corpus commun de français langue étrangère pour une étude phonétique des productions de locuteurs de langues maternelles plurielles. *Cahiers de Praxématique* 54-55, 73-86.

LÉON, P. R., LÉON, M. (2007). *La prononciation du français*. Paris: Armand Colin.

MADDIESON, I., WRIGHT, R. (1996). Small vowel systems and phonetic variability – evidence from Amis. *Austronesian Formal Linguistics Association* 3, 305-309.

MAEDA, S. (1989). Compensatory Articulation during Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes using an Articulatory Model. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Modelling*: Kluwer Academic Publishers.

MANUEL, S. Y., KRAKOW, R. A. (1984). Universal and language particular aspects of vowel-to-vowel coarticulation. *Haskins Status Report on Speech Research* 77-78, 69-78.

MILLER, L. J. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the Study of Speech*. NJ: Erlbaum, 39-73.

NOVAKOVA, S. (à paraître). La phonétique en cours de FLE : l'actualité tchèque. In M. Bořek-Dohalská, K. Suková Vychopňová (Eds.), *Didactique de la phonétique et phonétique en didactique du FLE*. Praha : Karolinum.

PAILLEREAU, N. (2015). *Perception et production des voyelles orales du français par des futures enseignantes tchèques de Français Langue Etrangère (FLE)*. Thèse non publiée, Sorbonne-Nouvelle, Paris.

PAILLEREAU, N. (à paraîtreA). Do isolated vowels represent vowel targets in French? An acoustic study on coarticulation. Actes du *5e Congrès Mondial de Linguistique Française*, Tours, France.

PAILLEREAU N. (à paraîtreB). "Identical" Vowels in L1 and L2? Criteria and Implications for L2 Phonetic Teaching/Learning. *EUROSLA Yearbook 2016*. Amsterdam: John Benjamins.

PUSTKA, E. (2011). L'accent méridional : représentations, attitudes et perceptions toulousaines et parisiennes. *Varia* 69, 117-152.

SKARNITZL, R., VOLÍN, J. (2012). Referenční hodnoty vokalických formantů pro mladé dospělé mluvčí standardní češtiny. *Akustické listy* 18, 7-11.

STEINLEN, A. K. (2005). *The influence of consonants on native and non-native vowel production. A cross-linguistic study*. Tübingen: GNV.

STEVENS, K. N., HOUSE, A. S. (1963). Perturbation of Vowel Articulations by Consonantal Context: An Acoustical Study. *Journal of Speech and Hearing Research* 6(2), 111-128.

STEVENS, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics* 17, 3-46.

STEVENS, K. N. (1998). Acoustic Phonetics. Cambridge: MIT Press.

ŠIMACKOVA, S., PODLIPSKÝ, V. J., CHLADKOVA, K. (2012). Czech spoken in Bohemia and Moravia. *Journal of the International Phonetic Association* 42(2), 225-232.

VAISSIÈRE, J. (1985). The use of allophonic variations of /a/ in automatic continuous speech recognition of French. *Journal of the Acoustical Society of America* 77(1), S12-S12.

VAISSIÈRE, J. (2006). *La phonétique*. Paris: Presses Universitaires de France.

VAISSIÈRE, J. (2009). Articulatory modeling and the definition of acoustic-perceptual targets for reference vowels. *The Chineses Phonetics Journal* 2, 22-33.

WIOLAND, F. (2005). *La vie sociale des sons du français*. Paris: L'Harmattan.

Pratique de la lecture en thaï et hindi en L2 : classification automatique de textes par progression lexicale

Jennifer Lewis-Wong[1, 2] Satenik Mkhitaryan[1]
(1) Équipe de Recherche Textes, Informatique, Multilinguisme (ERTIM - EA 2520), INALCO, 2 rue de Lille, 75343 Paris Cedex 07, France
(2) Langues et Civilisations à Tradition Orale - CNRS / Paris III / INALCO (LACITO - UMR 7107), 7 rue Guy Môquet (bât. D), 94801 Villejuif Cedex, France
jennifer.wong@inalco.fr, satenik.mkhitaryan@inalco.fr

Résumé

Cet article a pour objet la création automatique de ressources pour l'apprentissage de langues étrangères peu enseignées et peu dotées en matériels pédagogiques à partir de textes authentiques. Il s'inspire du travail de Ghadirian (2002) et son logiciel *TextLadder*, une application qui classifie les textes d'un corpus selon un ordre qui maximise la facilité de lecture pour l'apprenant, en calculant la similarité lexicale entre les textes. La classification automatique de textes par progression lexicale constitue une méthode intéressante pour proposer une séquence de textes appropriée au niveau d'un lecteur en L2, aussi bien pour proposer des textes à des lecteurs autonomes que pour la création de matériels pédagogiques destinés à être utilisés en classe. Cette méthode est spécialement bien adaptée à la classification de textes qui portent sur une thématique particulière.

Abstract

Text classification by lexical progression for L2 reading practice in Thai and Hindi

This article looks at the creation of teaching and learning resources for less commonly taught languages from unsimplified texts. The inspiration for this study comes from Ghadirian (2002) and the associated computer program TextLadder. *The program classifies a series of texts by their lexical similarity, introducing target vocabulary incrementally and thus making reading easier for the learner. This kind of automated text sequencing can be used to select sequences of texts appropriate to the level of lexical competence of the L2 reader, whether for independent readers or for creating teaching material for classroom use. The method is particularly suitable for classifying texts with a similar topic or theme.*

Mots-clés : ALAO, aide à la lecture, thaï langue étrangère, hindi langue étrangère, lisibilité, TextLadder.

Keywords: CALL, Reading aides, TFL, HFL, Readability, TextLadder.

1 Introduction

La pratique autonome et précoce de la lecture par un apprenant de langue étrangère (L2) présente maints avantages, notamment une exposition accrue à la langue étudiée et un réel plaisir.

Aujourd'hui, le web propose, pour d'innombrables langues, quantité de textes de tout genre, thème et taille. Il est ainsi susceptible de répondre aux besoins de lecture d'apprenants, même de langues, moins enseignées. Cette abondance de textes peut pourtant s'avérer être un handicap, voire un cauchemar, pour des apprenants de niveau initial ou intermédiaire, incapables de trouver des textes adaptés à leur niveau de compétence, qu'elle soit grammaticale ou lexicale.

Nous présentons dans cet article un dispositif automatique d'aide à la lecture, proposant un parcours de lecture au sein d'un corpus de textes optimisant l'acquisition de vocabulaire nouveau par le lecteur. Ce dispositif, inspiré de Ghadirian (2002), a été appliqué à deux langues fort diverses tant du point de vue linguistique (isolante / flexionnelle) que de leur système d'écriture (alpha-syllabaires sans / avec espaces), le thaï et le hindi, par ailleurs peu dotées en ressources pédagogiques.

Nous ferons tout d'abord, au chapitre 2, un rapide état de l'art concernant d'une part la notion de lisibilité des textes, c'est-à-dire d'évaluation de leur niveau de difficulté pour un apprenant, concernant d'autre part des études analogues portant aussi sur la classification de textes selon des critères lexicaux. Au chapitre 3 nous présenterons notre méthodologie : la constitution de listes lexicales de vocables, le choix des corpus de textes thaï et hindi et leurs prétraitements spécifiques, puis les principes de la classification et de création d'un parcours de lecture. Dans le chapitre 4, nous rendrons compte des tests de classification des deux corpus et des résultats obtenus. Enfin, le chapitre 5 proposera un regard critique sur l'étude et diverses pistes d'amélioration.

2 Travaux antérieurs

2.1 Sur la lisibilité des textes

Si les bienfaits de la lecture personnelle régulière sur l'acquisition d'une langue étrangère, surtout en ce qui concerne l'élargissement du vocabulaire, sont bien établis (Krashen, 2004), l'abondance de textes ne garantit pas que l'apprenant puisse trouver aisément des textes adaptés à son niveau. En effet, l'apprenant peut se démotiver face à un texte qui ne correspond pas à son niveau de connaissances lexicales. Pour les enseignants de langues peu enseignées et peu dotées en matériel pédagogique, choisir des textes qui correspondent au(x) niveau(x) de ses étudiants est également un défi. Selon Liu et Nation (1985) le lecteur en L2 doit connaître 95 % des mots d'un texte avant de pouvoir déduire le sens des mots inconnus. C'est ce qu'on nomme la couverture textuelle.

En règle générale, les chercheurs en lisibilité ont recours à deux types de stratégies pour évaluer la difficulté des textes. La première est l'élaboration de formules de lisibilité qui s'appuient sur des mesures de caractéristiques superficielles des textes, la deuxième stratégie est le développement de modèles statistiques plus complexes, basés sur des corpus de textes dont la difficulté a déjà été mesurée, comme les textes des manuels scolaires.

Ces méthodes sont peu adaptées aux langues peu enseignées en L2. En premier lieu, la plupart des travaux sur la lisibilité mesurent la difficulté des textes pour les locuteurs natifs. Or, François (2011) a montré que la lisibilité d'un texte pour un lecteur en L1 n'est pas la même que pour un lecteur en L2. Heilman et al. (2007, voir plus loin) ont trouvé que la difficulté grammaticale est plus décisive dans la lisibilité de textes en L2 qu'en L1. Ceci serait dû au fait que l'apprentissage du vocabulaire et l'apprentissage de la grammaire se déroulent en même temps en L2, alors que l'acquisition du vocabulaire continue après l'acquisition de la grammaire en L1.

Il faut donc développer de nouvelles formules, ou créer des modèles basés sur des corpus de textes destinés aux lecteurs en L2 déjà classifiés par niveau de difficulté. La deuxième difficulté est que ces méthodes sont développées sur une langue spécifique (jusqu'alors, la recherche sur la lisibilité en L2 s'est concentrée surtout sur l'anglais[1]) et ne sont pas nécessairement adaptées ou facilement adaptables à d'autres langues, bien que parfois une formule de lisibilité puisse donner des résultats satisfaisants pour des langues complètement différentes[2]. Le développement de modèles de lisibilité spécifiques à l'apprentissage en L2 des langues peu outillées et avec peu d'apprenants peut s'avérer impossible pour des raisons pratiques de temps et de ressources.

Le logiciel *TextLadder*, de Ghadirian (2002), développé pour des apprenants d'anglais langue étrangère, a l'avantage de ne pas dépendre de ces méthodes de lisibilité. Le dispositif sélectionne une séquence de textes à partir d'un corpus et les agence dans un ordre qui optimise la facilité de lecture, en prenant en compte une liste de vocabulaire connu et une liste de vocabulaire « cible » que le lecteur souhaiterait acquérir. Au lieu d'attribuer aux textes une note de difficulté, la classification est relative et plus fine ; le dispositif peut être paramétré pour prendre en compte les connaissances lexicales réelles de l'apprenant.

2.2 Applications de classification textuelle selon le lexique

REAP[3] est un système d'aide pédagogique pour des professeurs d'anglais L1 ou L2. Il a la particularité de prendre en compte les connaissances spécifiques de l'étudiant et son niveau d'études dans le choix des textes, des informations qui permettent la personnalisation du parcours de lecture. *REAP* a été porté en portugais et en français (Marujo et al., 2009).

La classification des textes par niveau de difficulté utilise une stratégie hybride, combinant modèles de langue basés sur la statistique lexicale et une classification par difficulté grammaticale modélisée sur des constructions grammaticales trouvées dans des manuels d'anglais langue étrangère de niveaux différents. La recherche sur laquelle *REAP* est basée (Heilman et al., 2007) a démontré que les modèles de langues basés sur la statistique lexicale sont plus efficaces que l'approche qui utilise la difficulté grammaticale seule, mais une combinaison des deux est encore plus précise à la fois pour des corpus L1 et L2.

1 Pour un état de l'art de la recherche sur la lisibilité en L2, lire François (2011).

2 Das et Roychudhury (2004; 2006), cités par Islam (2012) notent que l'indice de lisibilité de l'anglais Flesch–Kincaid donne des résultats satisfaisants pour le bengali, par exemple.

3 Le projet *REAP* est sous-titré *Reader-Specific Lexical Practice for Improved Reading Comprehension*. Voir Brown & Eskenazi (2004).

TextLadder est un logiciel conçu par Ghadirian (2002) qui classe des textes en anglais selon un ordre de lecture donné afin de faciliter l'acquisition du vocabulaire de manière progressive par la répétition lexicale. Le logiciel permet à l'utilisateur d'entrer son propre corpus de textes; une version web sur le site *ReadingEnglish* propose un corpus des textes d'anglais simplifié tirés du site de la Voice of America (VOA) classifiés de la même manière que *TextLadder*. Le site du projet taïwanais d'apprentissage Internet *Candle* inclut un module de pratique de la lecture appelé ***TextGrader***, conçu par Huang et Liou (2007) et basé sur le travail de Ghadirian (2002).

Les textes sont classifiés à l'aide de listes lexicales compilées pour l'anglais par d'autres chercheurs, la *GSL* (*General Service List*) de West (1953) et l'*UWL* (*University Word List*) de Xue et Nation (1984). Ces listes représentent respectivement le vocabulaire le plus fréquent de l'anglais et le vocabulaire supplémentaire le plus fréquent des textes académiques. Notons qu'il ne s'agit pas simplement de regrouper les unités lexicales en lemmes (différentes formes d'une même unité lexicale), mais de regrouper les unités lexicales apparentées dans des *familles de mots* (des unités lexicales qui ont la même racine et une ressemblance sémantique). À ces listes, Huang et Liou ajoutent la liste HSF (High School Frequency Word List), utilisée dans l'élaboration des manuels scolaires à Taïwan, et leur propre liste spécifique au corpus choisi. Ghadirian utilise aussi une liste de base de vocabulaire spécifique à son corpus, la *VOA Special English Word List*.

En premier lieu, ces listes sont utilisées pour filtrer les textes appropriés au lecteur. Seuls les textes couverts à au moins 95 % par ces listes sont retenus, suivant le principe déjà mentionné que le lecteur en L2 doit connaître 95 % des mots d'un texte avant de pouvoir déduire le sens des mots inconnus (Liu et Nation 1985). Ensuite sont créées deux listes : une liste du vocabulaire cible et l'autre de vocabulaire connu. Ghadirian n'utilise comme liste de vocabulaire connu que du vocabulaire connu des débutants : une partie de la GSL (les 176 premières familles de mots) augmentée avec du vocabulaire de base. Huang et Liou utilisent la totalité de la GSL plus la liste HSF comme liste de vocabulaire connu. Les autres listes servent à constituer la liste de vocabulaire cible dans chaque cas.

Après élimination de textes qui ne sont pas couverts à 95 % par les listes lexicales initiales, le dispositif choisit le texte le plus facile. C'est le texte avec le plus grand nombre de mots de la liste de vocabulaire connu et le moins de mots de la liste de vocabulaire cible. Ce texte devient le premier texte dans la séquence de textes à lire. Le vocabulaire cible identifié dans ce premier texte est ensuite rajouté à la liste du vocabulaire connu. Seul le vocabulaire de la liste de vocabulaire cible est rajouté, le vocabulaire inconnu qui ne figure pas dans la liste de vocabulaire cible n'est pas pris en compte. Cette nouvelle liste est utilisée pour identifier le deuxième texte de la séquence de lecture et ainsi de suite. Le système de *TextGrader* diffère légèrement de *TextLadder* dans la mesure où Huang et Liou ont conçu un algorithme qui favorise aussi la répétition de vocabulaire cible. Le vocabulaire cible rencontré n'est pas mis directement dans la liste de vocabulaire connu, mais dans une troisième liste, de vocabulaire cible exposé, le processus de classification favorisant le vocabulaire de cette liste.

Les textes de *TextGrader* sont présentés avec le vocabulaire cible en surbrillance, avec variation de la couleur de surbrillance selon qu'il s'agit de la première occurrence d'un mot ou qu'il s'agit d'une répétition. Les deux systèmes ont intégré un dictionnaire, *TextLadder* permettant au lecteur de rechercher la définition des tous les mots du texte, alors que *TextGrader* ne glose que le vocabulaire cible.

3 Méthodologie

Comme nous avons mentionné plus haut, cette étude s'inspire des travaux de Ghadirian (2002), l'objet est de créer un parcours de lecture dans un corpus de textes, proposant à chaque étape le texte optimisant au mieux l'acquisition du vocabulaire du corpus. Nous présentons, dans ce qui suit, la création des différentes listes, les corpus, les prétraitements spécifiques à chaque langue, ainsi que les étapes de l'implémentation de la méthode.

3.1 Listes lexicales

Nous avons pour objectif l'acquisition du vocabulaire le plus utile, c'est-à-dire le lexique le plus fréquent et le vocabulaire spécifique au corpus. Nous avons trouvé que le seuil de lexique de haute fréquence se trouve à environ 5000 mots. Ce sont ces mots que nous avons utilisés pour former nos listes de vocabulaire connu et cible, complétées par d'autres listes de vocabulaire de langue parlée que nous avons jugé connu d'étudiants de niveau intermédiaire. Nous avons aussi utilisé notre propre intuition en tant qu'apprenantes de ces langues pour situer la frontière entre le vocabulaire connu et cible en nous référant aux manuels d'apprentissage de thaï et de hindi. Au-delà de ces 5000 mille mots se trouve la majorité du lexique d'une langue qui est de basse fréquence. Notre système sélectionne du vocabulaire de basse fréquence qui est fréquent dans le corpus, afin de créer une liste de vocabulaire spécifique au corpus utile pour l'apprenant. Cette liste de vocabulaire spécifique est rajoutée à la liste de vocabulaire cible.

Par souci de cohérence avec les textes de nos corpus, nous avons choisi d'utiliser une liste de fréquence basée sur un corpus généré avec la méthode *Corpus Factory* (Kilgarriff et al., 2010)[4] sur le site de *Sketch Engine*[5] pour notre projet. Il se trouve que l'outil de segmentation du thaï et le lemmatiseur du hindi utilisés par *Sketch Engine* (*SWATH*, Hindi POS Tagger, voir ci-dessous) sont les mêmes que nous utilisons pour segmenter le corpus thaï et lemmatiser le corpus hindi.

Sketch Engine est un outil de création de corpus à partir du web, disponible pour un grand nombre de langues, dont 36 possèdent déjà un corpus de référence. Le corpus thaï *ThaiWaC* fournit par *Sketch Engine*, contient environ 108 M tokens et le corpus hindi *HiWaC* en contient environ 65 M. A partir de ces corpus, nous avons obtenu des listes de fréquence triées par fréquence lexicale, chaque élément accompagné du nombre de ses occurrences dans le corpus de référence.

Lors de nos essais sur le corpus de presse thaï, nous avons trouvé que nos listes de vocabulaire de haute fréquence ne couvrent en moyenne que 72 % du vocabulaire des textes de nos corpus, très loin des 95 % recommandé par Liu et Nation (1985) pour la déduction du sens des mots inconnus. Afin de couvrir suffisamment le vocabulaire de basse fréquence pour atteindre les 95 %, notre dispositif complète la liste de vocabulaire cible avec une liste de vocabulaire spécifique au corpus de textes à classifier. Créée à la volée, cette liste, en combinaison avec la liste de haute fréquence, couvre plus aisément 95% des textes. Ceci diffère de la stratégie employée par Ghadirian (2002) et Huang et

[4] La méthode *Corpus Factory* détaillée dans Kilgarriff et al. (2010) consiste à télécharger une sauvegarde Wikimédia pour une langue donnée afin de créer un corpus à partir de Wikipédia et générer une liste de fréquence lexicale. Les éléments de moyenne fréquence sont utilisés pour interroger des moteurs de recherche et récupérer des pages web, qui sont nettoyées de tout balisage et publicité, avant d'être filtrées pour créer un corpus « propre ».

[5] https://www.sketchengine.co.uk/

Liou (2007) dont les systèmes de classification ont recours à des listes préfabriquées spécifiques non seulement au type de texte, mais spécifiques au corpus. Notre stratégie permet d'éviter les inconvénients de listes spécifiques figées : l'élaboration de telles listes de vocabulaire spécifique exige un corpus de textes représentatif (qui n'est pas toujours disponible pour des langues peu dotées) et limite le choix de textes que le système peut classifier.

3.2 Corpus

Pour le thaï, le corpus principal que nous avons utilisé pour nos tests est composé d'articles apparus entre 2013 et 2014 dans un journal quotidien populaire à grand tirage, *ThaiRath*, divisé en huit rubriques. Le corpus hindi contient des articles de presse de l'année 2013 apparus sur le site de NDTV (New Delhi Television Limited) dans la rubrique « India ».

3.3 Prétraitements de textes

Les prétraitements à effectuer sur les textes sont nécessairement dépendants des particularités d'une langue et de son système d'écriture. Comme mentionné précédemment, il est préférable que les prétraitements effectués sur les textes soient les mêmes que ceux utilisés dans la création des listes.

3.3.1 En thaï

Comme tous les systèmes d'écriture de l'Asie du Sud-est qui utilisent des alphasyllabaires, ainsi que le chinois et le japonais, le thaï s'écrit en continu sans séparation entre les mots (*scriptio continua*). Entre ces systèmes, le thaï moderne a la particularité de ne pas posséder de signe de ponctuation spécifique à la délimitation de la phrase. Il n'y a pas non plus de distinction majuscule/minuscule qui permettrait d'identifier facilement le début de phrase et les entités nommées (c'est-à-dire les noms de personnes, les toponymes et les noms d'organisations). Pour le traitement automatique de ce style de texte, une étape de division de l'énoncé en unités lexicales est nécessaire avant tout autre traitement, par le biais d'un outil de segmentation. Nous utilisons l'outil de segmentation nommé *SWATH* (*Smart Word Analysis for THai*) de Meknavin et al. (1997). Cet outil sépare les mots d'éventuels signes de ponctuation thaï, comme le symbole de répétition. *SWATH* améliore la segmentation des mots inconnus avec une analyse linguistique qui prend en compte l'environnement des mots, cherchant des mots de contexte et des collocations pour déterminer la segmentation la plus probable.

วัน|หนึ่ง|_|เขา|ตื่น|ลืมตา|ขึ้น|มา|จาก|ความ|ฝัน|อัน|สับสน|_|รู้สึก|อ่อน|_|เพลีย|จาก|การ|พักผ่อน|ไม่|เพียงพอ|_|เขา|เดิน|โซเซ|ไป|ชนก|อง|หนังสือ|ใน|ห้องรับแขก|ที่ตั้ง|ไว้|สูง|จนเป็น|กำแพง|กระดาษ|_|หนังสือ|เล่ม|ใหญ่|เกือบ|สิบ|เล่ม|หล่น|ลง|มา|ทับ|เท้า|ทำเอา|เขา|ร้อง|โอดโอย|_|แล้ว|เซ|ไป|เตะ|ซีดี|ที่|กอง|ไว้|อีก|มุม|ห้อง|จน|กระจัดกระจาย|_|ต้อง|เดิน|กะเผลก|_|ๆ|_|ผ่าน|ห้องหับ|อัน|มัวซั่ว|ไป|เข้า|ห้องน้ำ|ล้างหน้าล้างตา

FIGURE 1: Exemple de segmentation résultant d'un traitement avec SWATH (Le trait vertical, |, sépare les tokens et le tiret bas, _, représente l'espace dans le texte avant la segmentation)

Les erreurs de segmentation proviennent souvent de la présence d'entités nommées, mais il faut comprendre aussi que la segmentation est ambiguë dans le cas de mots composés, dans ces cas seul le contexte permet de décider définitivement. Compte tenu du fait qu'il s'agit d'une langue sans flexions, une étape de lemmatisation n'est pas nécessaire dans le prétraitement de textes en thaï.

3.3.2 En hindi

Le hindi utilise le système d'écriture devanāgarī (देवनागरी), un système également alphasyllabique, qui s'écrit de gauche à droite avec des espaces entre les mots. La devanāgarī dispose de ses propres symboles pour représenter les chiffres (० १ २ ३ ४ ५ ६ ७ ८ ९), mais les chiffres arabes sont de plus en plus souvent employés. Le hindi, comme la thaï, ne fait pas de distinction entre majuscules et minuscules ce qui rend opaques les entités nommées. La devanagari utilise les signes de ponctuation de l'alphabet latin, sauf la fin de phrase qui est marquée par une barre verticale « । » propre à cette écriture).

Le hindi est une langue à flexion nominale, adjectivale et verbale, donc une phase de lemmatisation est nécessaire pour le corpus hindi. Nous avons utilisé le lemmatiseur intégré à l'étiqueteur morphosyntaxique de Reddy et Sharoff (2011). La lemmatisation résultante n'est pas dépourvue d'erreurs, mais les erreurs sont cohérentes avec nos listes de vocabulaire qui ont été lemmatisées avec le même outil.

3.4 Principes de classification

À l'instar du logiciel *TextLadder* de Ghadirian (2002), notre dispositif dispose d'une liste de vocabulaire connu et d'une liste de vocabulaire cible, c'est-à-dire le vocabulaire à acquérir. Après des prétraitements spécifiques à chaque langue, le classement automatique des textes suit les étapes suivantes :

1. Sélection des textes du corpus par la longueur (par défaut entre 300 et 1500 mots).

2. Création d'une liste de vocabulaire spécifique au corpus. Cette liste comprend tous les éléments qui ne figurent pas dans les listes initiales, qui ont une fréquence suffisante et représentative (nous avons choisi respectivement huit occurrences et cinq textes).

3. Ajout de la liste du vocabulaire spécifique au corpus à la liste du vocabulaire cible.

4. Sélection des textes qui sont couverts à 95% par l'ensemble des listes. Sont considérés comme appartenant au vocabulaire connu tous les mots en lettres latines et les logogrammes (tels que les chiffres arabes, les symboles monétaires, etc.).

5. Choix du texte le plus accessible d'un point de vue lexical. Deux critères de choix sont possibles : soit la maximisation du vocabulaire connu, soit la maximisation de la couverture textuelle.

6. Ajout du texte à la séquence de lecture.

7. Ajout du vocabulaire cible du texte sélectionné à la liste de vocabulaire connu.

Répétition des étapes 5 à 7, choisissant comme texte suivant le texte le plus facile qui n'a pas encore été sélectionné.

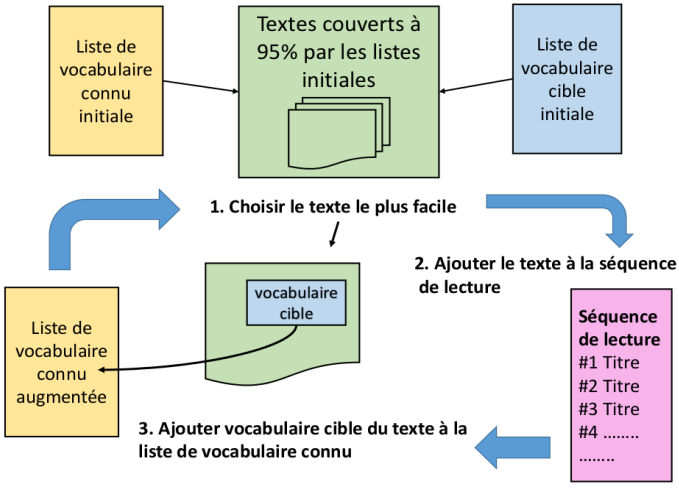


FIGURE 2: Fonctionnement de base

Le vocabulaire cible des textes est balisé pour informer le lecteur s'il s'agit de la première occurrence du vocable, ou d'une répétition. La première occurrence du vocable dans la séquence de textes est soulignée en vert, les occurrences suivantes sont soulignées avec de différentes nuances de bleu, qui s'éclaircissent progressivement au fur et à mesure des répétitions

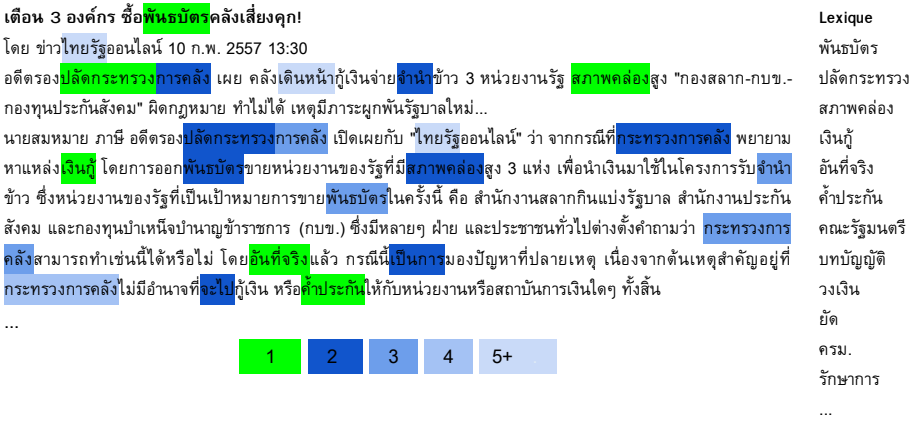


FIGURE 3: Balisage du vocabulaire cible d'un texte du corpus *ThaiRath2013* rubrique *Business* issu d'une séquence de lecture

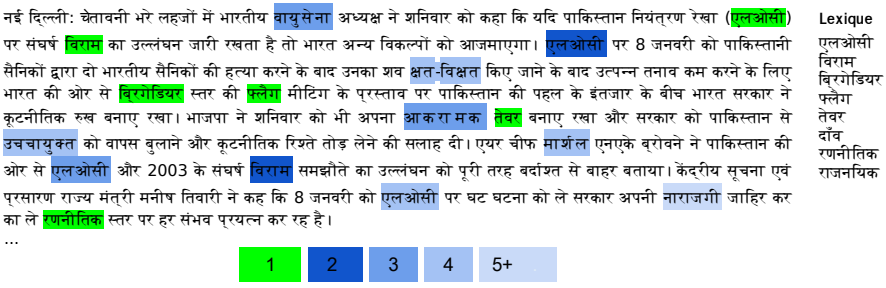


FIGURE 4: Balisage du vocabulaire cible d'un texte du corpus *NDTV Hindi* rubrique *India* issu d'une séquence de lecture

4 Tests sur corpus et résultats

4.1 La segmentation du thaï

Tout comme Jean (2009), nous avons trouvé beaucoup d'instances de sursegmentation d'entités nommées dans le corpus thaï. L'outil de segmentation, qui n'a pas ces mots dans son dictionnaire, les divise en syllabes. Notre dispositif de création de liste de vocabulaire spécifique au corpus destinée à être ajoutée à la liste de vocabulaire cible va donc introduire soit des mots inexistants dans la langue, soit risquer d'y introduire des mots de très faible fréquence dans des cas d'homographie entre d'autres mots et ces syllabes. Ce problème de segmentation ne concerne pas uniquement les noms propres, mais aussi les mots anglais écrits en thaï et les mots composés thaï-anglais. Nous avons constaté un grand nombre de ces mots dans les textes de presse, certains fréquemment utilisés en thaï peuvent se considérer comme des emprunts intégrés et d'autres, clairement considérés comme des mots étrangers à la langue, n'existant dans aucun dictionnaire de thaï.

Ayant constaté que beaucoup d'entités nommées sont entourées d'espaces typographiques, nous avons introduit un dispositif d'amélioration de leur segmentation avant l'étape de la création de la liste de vocabulaire spécifique au corpus, à l'aide d'une liste d'entités nommées extraite d'une sauvegarde des entêtes des pages Wikipédia en thaï. Ceci a eu l'effet d'améliorer sensiblement la segmentation des noms de personnes et de lieux avec très peu de cas de sous-segmentation.

4.2 Influence de types de tri

Nous avons pris un échantillon de 300 textes thaï pris au hasard de notre corpus d'articles de presse *ThaiRath*, rubrique *Business* pour un premier aperçu de la classification de textes en séquence de lecture. Le tri par quantité de vocabulaire connu, qui favorise le texte avec le plus grand nombre de mots connus est comparé avec le tri par couverture textuelle, qui lui favorise le texte avec le plus grand pourcentage de couverture textuelle de la liste de vocabulaire connu. Pour cette dernière, si plusieurs textes ont le même pourcentage de couverture textuelle, le dispositif choisit le prochain texte à mettre sur la séquence de lecture selon les critères suivants : si le pourcentage de couverture

est en dessous de 95 %, il choisit celui qui a le moins de vocabulaire cible, mais si la couverture textuelle dépasse 95 %, le texte avec le plus de vocabulaire cible est choisi. Ceci permet d'augmenter la couverture textuelle de vocabulaire connu et de répartir le nouveau vocabulaire plus équitablement.

Les résultats de ces deux types de tri sont illustrés par les figures 5 et 6. La figure 6 montre l'évolution de la taille du nouveau vocabulaire sur la séquence de lecture, la figure 5 représente la couverture textuelle par la liste de vocabulaire connu ; pour chaque graphique l'axe horizontal représente la séquence de lecture, par numéro de texte sur la séquence.

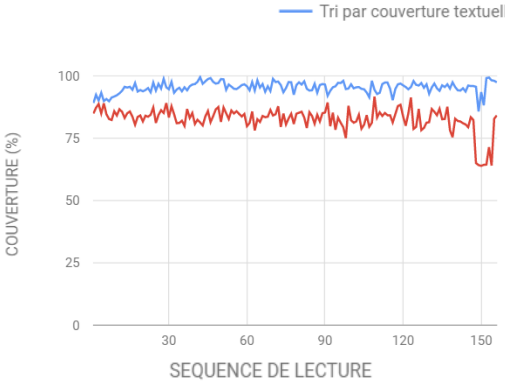


FIGURE 5: Couverture textuelle par liste de vocabulaire connu

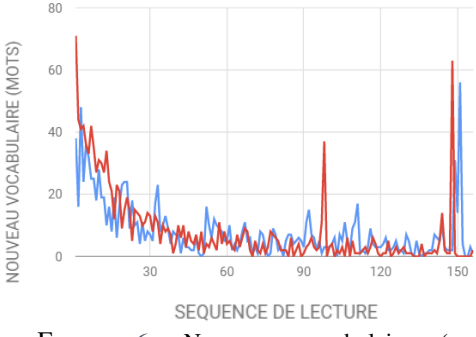


FIGURE 6: Nouveau vocabulaire (corpus ThaiRath rubrique Business de 300 textes)

On constate que l'évolution de la taille du nouveau vocabulaire est moins abrupte pour le tri par couverture textuelle que le tri par quantité de vocabulaire. Dans le cas du tri par couverture textuelle, le premier texte a un vocabulaire de 38 éléments alors que le tri par taille du vocabulaire mettait un texte d'un vocabulaire de 71 éléments en première position. Le nombre de textes sans nouveau vocabulaire a diminué légèrement (8 textes, contre 11 avec le tri par taille) et le nombre de textes avec un seul vocable nouveau a aussi diminué (le nombre de textes est passé de 19 à 15). Le tri par couverture textuelle permet également d'assurer une meilleure couverture textuelle, pour dépasser le seuil de 95% à partir du texte numéro 13. 106 des 156 textes ont une couverture textuelle d'au moins 95% par la liste de vocabulaire connu.

4.3 Homogénéité du corpus

Pour tester un classement de textes plus hétérogènes, nous avons testé notre système sur un corpus de 300 articles en thaï de *ThaiRath* pris dans diverses rubriques (*Business*, *Divertissement*, *Sport*, *Lifestyle*, *National* et *International*).

Nous constatons d'abord que des 300 textes, seuls 20 sont couverts à 95 % avec les listes de vocabulaire initiales. Ceci est dû au fait que la liste de vocabulaire spécifique créée automatiquement pour un corpus hétérogène est trop petite. La liste initiale du vocabulaire spécifique à notre corpus de textes de sources mixtes (801 vocables) est réduite de moitié par les critères de fréquence minimum et nombre de textes minimum, alors que la liste initiale du corpus constitué de 300 articles

tirés exclusivement de la rubrique *Business*, initialement de 1115 vocables, atteint toujours 723 vocables après application de ces critères. Pour examiner la répétition de vocabulaire, nous avons réduit l'exigence de couverture des listes initiales à 85 % pour disposer de plus d'articles. Nous avons constaté une plus grande proportion d'hapax dans le corpus hétérogène, mais il faut noter qu'il s'agit en grande partie d'entités nommées.

4.4 Application au corpus hindi

Nous avons choisi un sous-corpus de 1030 articles du corpus *NDTV Hindi*. Afin d'observer l'impact de la lemmatisation des textes, nous avons testé d'abord avec un corpus et des listes de vocabulaires non lemmatisés. Le système a sélectionné 160 textes couverts à plus de 95 %. Après la lemmatisation du corpus et des listes de vocabulaires, le nombre de textes sélectionnés a augmenté jusqu'à 183. Ceci montre que la lemmatisation n'a pas eu beaucoup d'effets sur les résultats, car le lemmatiseur choisi ne lemmatise qu'un faible pourcentage de mots et, de surcroît, fait de nombreuses erreurs. Néanmoins nous avons continué les tests avec le corpus lemmatisé, qui est plus intéressant qualitativement en termes acquisition de vocabulaire.

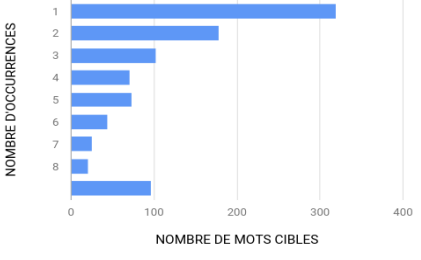


FIGURE 7: Distribution du vocabulaire cible

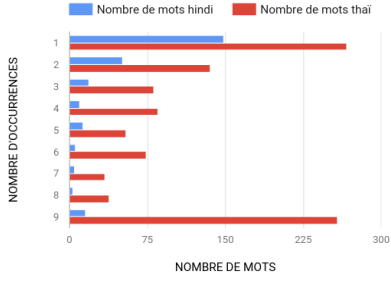


FIGURE 8: Répétitions du vocabulaire cible

La figure 7 montre un grand nombre d'hapax. Ceci est dû en grande partie à la présence des entités nommées, notamment des toponymes, et de mots anglais qui pourraient être considérés comme du vocabulaire connu. L'utilisation des mots anglais en hindi peut être aléatoire donc il est difficile de les recenser. Nous avons rencontré ce phénomène en thaï aussi, dans une moindre mesure. Un détecteur de mots anglais translittérés dans les langues traitées améliorerait la couverture des listes de vocabulaire connu. Le corpus thaï *Business* serait plus homogène que le corpus hindi *India* donc il contiendrait plus de répétitions (figure 8).

5 Discussion

La présente étude possède un certain nombre de limites méthodologiques qu'il convient de prendre en compte dans une discussion sur les résultats que nous avons obtenus. D'abord, nous avons observé une forte corrélation entre les performances de la méthode (à la fois sur le plan du nombre de répétitions du vocabulaire et du choix des textes) et la qualité des prétraitements effectués sur nos

corpus, en l'occurrence la segmentation du thaï et la lemmatisation du hindi. Nous prévoyons de tester d'autre lemmatiseurs du hindi pour tenter d'améliorer les résultats. D'autres facteurs ont aussi influé les performances de notre dispositif, tels que la présence d'entités nommées et des mots anglais, qui selon nous pourrait être considérés comme du vocabulaire acquis. Les résultats pourraient donc être améliorés par l'intégration de traitements tels que la détection automatique de mots anglais translittérés et des entités nommées, afin de proposer à l'utilisateur l'option de les considérer comme du vocabulaire connu.

Nous sommes conscientes que notre système reflète une vision assez simpliste du vocabulaire, car il ne prend pas en compte la polysémie des lexèmes, exagérée pour des lexèmes d'une langue isolante, comme le thaï, dépourvus de repères que donne la syntaxe. Ghadirian (2002) signale lui-même que sa méthode n'analyse pas les collocations, ou les verbes à particule (*phrasal verbs*). Nous ne tenons pas compte non plus du fait que l'apprenant peut acquérir les différents sens d'un lexème à des niveaux différents. En effet, il est difficile de situer les connaissances de l'apprenant sur un continuum de compétence lexicale, car tout dépend des connaissances extralinguistiques du monde réel de l'apprenant, et cela peut varier considérablement. Sur ce plan, les entités nommées et mots anglais écrits en thaï et en hindi nous ont donné beaucoup de matière à réflexion. L'efficacité du dispositif dépend de la capacité des listes de vocabulaire connu et cible de représenter les connaissances lexicales réelles de l'apprenant. Un dispositif de test de niveau avant, voire au cours de la lecture du parcours proposé est à l'étude.

Avant d'entreprendre des modifications possibles, il faudrait définir l'objectif de la lecture. S'il s'agit d'une lecture intensive destinée à l'apprentissage lexical ou syntaxique, ou une pratique extensive de la lecture pour la compréhension du contenu et le plaisir de découvrir du texte authentique. En effet, Ghadirian (2002) prévoit un usage de *TextLadder* pour la lecture intensive des textes avec une quantité importante de nouveau vocabulaire, alors que Huang et Liou (2007) ont utilisé cette méthode de classification de textes pour encourager la lecture extensive. Dans le cas de la lecture extensive, est-il nécessaire d'augmenter et cibler les expressions idiomatiques et les collocations ? Ne serait-il pas préférable de laisser le lecteur identifier les cas de polysémie ? À ces questions s'ajoute le fait que le dispositif ne prend en compte que l'aspect lexical dans le calcul de difficulté des textes et vise principalement l'acquisition lexicale.

6 Conclusion

Les résultats que nous avons obtenus nous permettent de démontrer qu'il est possible de créer des ressources pédagogiques pour la lecture en L2 de façon automatique avec peu d'intervention humaine, sans avoir recours à des modèles ou des formules de lisibilité. Ceci est particulièrement utile dans le traitement d'une langue peu enseignée, pour laquelle les corpus d'entraînement déjà classifiés par niveau de difficulté n'existent pas. En dépit des difficultés méthodologiques rencontrées, le dispositif est un moyen efficace de proposer des textes au niveau de connaissance lexicale de l'apprenant et le parcours de lecture créé permet une répétition de vocabulaire suffisante pour faciliter l'acquisition du vocabulaire nouveau avec un effort moindre pour le lecteur. La méthode marche mieux pour un corpus de thématique homogène, idéal pour la création de matériels pédagogiques pour l'enseignement de langues sur objectifs spécifiques.

Références

François T. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère.* Thèse de doctorat, Université Catholique de Louvain.

Ghadirian S. (2002). Providing controlled exposure to target vocabulary through the screening and arranging of texts. *Language Learning & Technology* 6(1), 147-164.

Heilman M., Collins-Thompson K., Callan J., Eskenazi M. (2007). Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. Actes de *The Human Language Technology Conference. Rochester, NY.*

Huang H-T., Liou H-C. (2007). Vocabulary Learning in an Automated Graded Reading Program. *Learning & Technology* 11(3), 64-82.

Islam Z., Mehler A., Rahman R. (2012). Text Readability Classification of Textbooks of a Low-Resource Language. Actes du *26th Pacific Asia Conference on Language, Information and Computation, 2012.*

Jean C. (2009). Le thaï. De la segmentation aux maux. *Lexicometrica. Numéro spécial - Explorations textométriques.* 2009.

Kilgarriff A., Reddy S., Pomikálek J., Avinesh PVS. (2010). A Corpus Factory for many languages. Actes de *LREC, Malta, 2010.*

Krashen S. (2004). *The Power of Reading: Insights from the Research*. Libraries Unlimited, Connecticut & London.

Liu N., Nation I. S. P. (1985). Factors affecting guessing vocabulary in context. *RELC Journal* 16(1), 33-42.

Meknavin S., Charoenpornsawat P., Kijsirikul B. (1997). Feature-based Thai Words Segmentation. *NLPRS, Incorporating SNLP-97.*

Marujo L., Lopes J., Mamede N., Trancoso I., Pino J., Eskenazi M., Baptista J., Viana C. (2009). Porting REAP to European Portuguese. Actes du *SLaTE Workshop on Speech and Language Technology in Education.*

Reddy S., Sharoff S. (2011). Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources. *Cross Lingual Information Access* 11.

West M. (1953). *A general service list of English words*. Londres : Longman, Green & Co.

Xue G., Nation I. S. P. (1984). A university word list. *Language Learning and Communication* 32, 215-219.

Un logiciel pour l'enseignement de la prosodie

Philippe Martin
LLF, UFRL, Université Paris Diderot Sorbonne Paris Cité
Place Paul Ricœur, 75013 Paris, France
philippe.martin@linguist.univ-paris-diderot.fr

RESUME

On présente un logiciel de visualisation en temps réel de la mélodie de la phrase, WinPitch LTL, dont la dernière version est optimisée pour une utilisation pédagogique. Reprenant la disposition traditionnelle modèle-imitation sur deux fenêtres séparées, ce logiciel est équipé de fonctions qui permettent de répondre aux insuffisances souvent constatées dans d'autres réalisations. Celles-ci portent essentiellement sur l'absence d'indications visuelles claires sur les courbes mélodiques qui rendraient compte des éléments pertinents au regard d'une approche théorique convaincante et explicite. Une fonction d'alignement automatique permet alors de reporter sur la courbe de l'apprenant les segments mélodiques jugés pertinents, conduisant ainsi celui-ci à juger de sa performance. D'autres fonctions du logiciel comme le play-back au ralenti et le morphing prosodique en renforcent les aspects pédagogiques.

ABSTRACT

A visualizer to teach sentence prosody
This paper presents a visualization software of sentence intonation operating in real time, WinPitch LTL, whose latest version is optimized for educational use. Taking the traditional model-imitation layout on two separate windows, this software has features that may meet the shortcomings often found in comparable realizations. These shortcomings mainly pertain to the absence of clear visual indications on the displayed melodic curves that would be related to some explicit theoretical approach. An automatic aligner maps model selected segments, deemed relevant, to the learner melodic segments, leading the learner to judge her/his performance. Other functions like slow playback and prosodic morphing reinforce the educational aspects of the software.

MOTS-CLES : intonation, prosodie, enseignement de l'oral, structure prosodique incrémentale.
KEYWORDS: intonation, prosody, language teaching, incremental prosodic structure.

1. Introduction

La visualisation de l'intonation à des fins pédagogiques, réalisé en temps réel ou différé, remonte au moins à 1964 (Vardanian, 1964). Beaucoup d'autres réalisations ont été proposées depuis (entre autres Léon et Martin, 1972 ; Abberton et Fourcin, 1975 ; Bot, 1980) avec des résultats pratiques discutables en ce qui concerne l'enseignement oral d'une langue seconde. Une des raisons expliquant leur possible inefficacité semble découler de l'absence de tout contexte théorique qui guiderait les apprenants, auxquels il était souvent demandé d'imiter globalement une courbe mélodique sans autre explication (James, 1976). De nombreux utilisateurs ont été rapidement découragés lorsque

l'acceptabilité de leurs réalisations n'a pas été évaluée par rapport à un objectif intonatif clair et un modèle phonologique ou phonétique explicite.

Pour être plus efficace, il apparait que la visualisation devrait être accompagnée d'une part d'explications précises quant aux réalisations prosodiques souhaitées, et de l'autre de la localisation automatique sur la courbe mélodique de l'apprenant des segments mélodiques admis comme pertinents. Cette pertinence, qu'elle soit phonologique ou phonétique, doit découler d'un modèle rendant compte des hauteurs ou des mouvements mélodiques spécifiques survenant sur telle ou telle syllabe, et non de caractéristiques globales de la courbe mélodique (durée, hauteur moyenne, etc.) (Martin, 1975, 2009). L'existence implicite d'un modèle est particulièrement importante compte tenu des variations acceptables de locuteurs natifs. Sinon, le système pourrait refuser une réalisation de l'apprenant, parfaitement acceptable pour un natif, mais dont le modèle et l'évaluation n'aurait pas tenu compte.

2. Des améliorations possibles

Les réalisations antérieures de visualiseur de mélodie à but d'enseignement de l'oral semblent donc, pour la plupart, souffrir de deux défauts : 1) le manque de modèle implicite pout guider l'apprenant dans ses réalisations prosodiques et 2) l'absence d'un système de localisation des segments considérés comme fautifs relativement au modèle dans la réalisation de l'apprenant. Ce dernier point doit permettre d'attirer l'attention sur des corrections ciblées et non globales. De plus, pour surmonter l'appréhension de certaines catégories d'apprenants, un tel système devrait pouvoir afficher, en plus de courbes mélodiques modèle et imitée, courbes relativement complexes qui peuvent désorienter les utilisateurs, des messages précis relativement aux erreurs constatées par le système. La visualisation deviendrait alors optionnelle.

WinPitch LTL, le logiciel présenté ici, a été spécialement conçu pour répondre à ces deux exigences en intégrant clairement sur les courbes mélodiques affichées des caractéristiques phonologiques jugées importantes par un expert auteur, lequel est libre d'élaborer des ensembles de phrases regroupées en leçons selon le modèle théorique choisi.

Les deux améliorations potentielles sont donc : a) un modèle explicite rendant compte des variations mélodiques (et donc de la structure prosodique des énoncés) dont les caractéristiques pertinentes sont affichées sur les courbes mélodiques et b) une fonction permettant de retrouver automatiquement les segments pertinents dans les réalisations de l'apprenant malgré les différences possibles de débit et de rythme par rapport au modèle.

3. Un modèle prosodique

L'ensemble du système étant ouvert grâce à de nombreuses fonctions auteur permettant d'élaborer des leçons d'exercices annotés au gré de l'instructeur, tout modèle prosodique peut convenir, pourvu qu'il soit aisément compréhensible par l'apprenant et que l'instructeur puisse en traduire les éléments pertinents sur les courbes mélodiques des énoncés modèles choisis. À côté des modèles phonologiques dominants, en particulier ceux dérivés de l'approche autosegmentale-métrique (Jun & Fougeron, 2002), on rappelle ici les grandes lignes d'un modèle théorique centré sur le caractère dynamique sur l'axe du temps des évènements prosodiques, tant du point de vue du locuteur que de celui de l'auditeur. Ce caractère dynamique implique entre autres que les évènements prosodiques, ont un caractère local et non global, et que leurs réalisations sous forme de contours mélodiques dépendent en général de leur contexte immédiat.

Du point de vue théorique choisi, il s'agit pour l'élaboration des exemples de marquer par différentes couleurs selon leur nature les contours mélodiques placés sur les syllabes accentuées et

singulièrement sur les voyelles accentuées. Ces syllabes sont considérées par hypothèse comme seules pertinentes pour l'indication de la structure prosodique de l'énoncé.

Le modèle de structure prosodique incrémentale est basé sur un principe clairement défini, l'indication d'une relation de dépendance "à droite" entre les événements prosodiques instanciés par des contours mélodiques placés sur les syllabes accentuées des groupes accentuels. Ces groupes se trouvent en français en position finale de syntagmes intonatifs. À partir d'un inventaire des classes de contours mélodiques en termes de relations de dépendance, une description simple de la structure prosodique émerge à partir du principe de contraste de pente.

Ce principe spécifie qu'un contour mélodique Cx indique une relation de dépendance envers un autre contour Cy "à droite" (i.e. apparaissant plus tard dans l'énoncé) par une pente mélodique de sens opposée. Ainsi un contour C1, souvent appelé de continuation majeure, indique une relation de dépendance envers un contour conclusif terminal déclaratif C0 par une variation mélodique montante, opposée à celle, descendante, du contour terminal. Ces relations de dépendance déterminent les regroupements successifs des unités prosodiques minimales que sont les groupes accentuels, c'est-à-dire les séquences de syllabes terminées par une syllabe accentuée (hors accent emphatique ou d'insistance).

On aboutit ainsi à une liste de contours mélodiques utilisés en français pour l'indication de la structure prosodique d'un énoncé.

Si C0, C1, C2 et Cn désignent ces contours mélodiques réalisés sur les syllabes (les voyelles) effectivement accentuées de l'énoncé, on a :

C0 contour descendant et bas ; C1 contour montant ; C2 contour descendant ; Cn contour neutralisé, de faible variation mélodique.

Pour attirer l'attention de l'apprenant sur leur pertinence, on peut surligner en couleurs différentes chacune de ces classes de contours. Par exemple, C0 rouge, C1 bleu, C2 marron, Cn noir, etc. au gré de l'auteur des exemples.

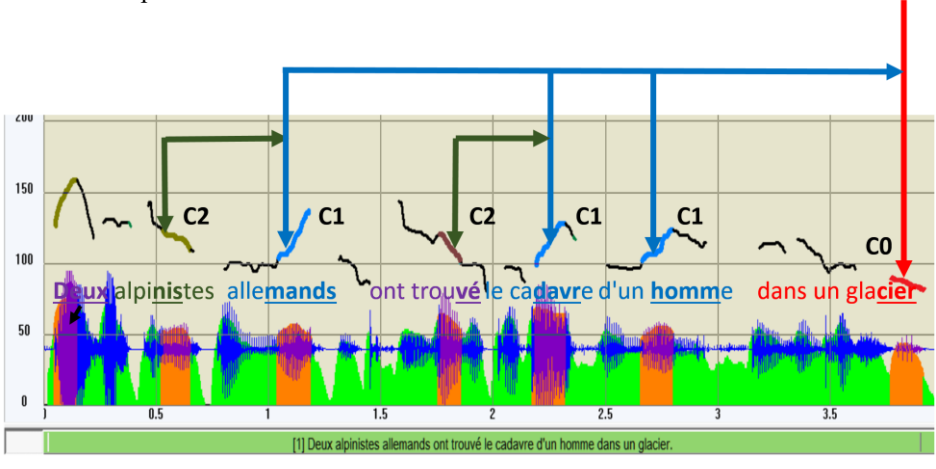


FIGURE 1 : Exemple d'un réseau de relations de dépendance indiquées par les flèches horizontales, chacun des contours pertinents du point de vue de l'approche théorique choisie par l'auteur étant colorée dans différentes couleurs (ici C0 rouge, C1 bleu, C2 marron). Les regroupements successifs des groupes accentuels sont indiqués par le parenthésage *[[deux alpi**nist**es C2][alle**mand**s C1]][[ont trou**vé** C2][le ca**davr**e C1]][d'un **homm**e C1][dans un gla**cier** C0]*

Les relations de dépendance indiquées par les contours sont résumées par l'expression Cn -> C2 -> C1 -> C0. Ces relations sont transitives, c'est-à-dire que Cn -> C2 implique aussi Cn -> C1 et Cn -> C0 (les flèches Cx -> Cy indiquent la dépendance de Cx par rapport à Cy). Une séquence C2 C0 est extrêmement rare et procède d'une marque d'insistance (Martin, 2009).

Les relations de dépendance opèrent « à droite » relativement au futur de la séquence de contours mélodiques. Ainsi un contour C1 de continuation majeure, en indiquant non seulement que l'énoncé n'est pas terminé, présuppose l'apparition d'un contour terminal C0 signalant la fin de l'énoncé et permettant à l'auditeur de rassembler les différents groupements de syntagmes pour accéder au sens de l'énoncé. L'existence de C1 dépend donc de l'apparition future d'un contour C0. Il en va de même pour le contour descendant C2, qui dépend de l'apparition future d'un contour C1, etc.

On notera que ce mécanisme d'indication de la structure prosodique par regroupements successifs de groupes accentuels va clairement à l'encontre du cadre théorique dominant basé sur la phonologie métrique-autosegmentale. En niant par hypothèse le rôle des accents mélodiques (pitch accents), cette approche ne peut évidemment rendre compte d'une quelconque interaction entre les contours mélodiques. Le cadre autosegmental-métrique ne propose en fait aucune explication quant à la fonction de la structure prosodique autre qu'actualiser la syntaxe, en se basant sur le concept du « bien-formé » emprunté à l'approche générative transformationnelle. Or il est relativement facile de démontrer que la structure prosodique n'est pas dérivée de la syntaxe (si ce n'est dans une certaine mesure dans l'oralisation de l'écrit), mais qu'au contraire elle la précède aussi bien pour le locuteur que pour l'auditeur (Martin, 2015).

4. Alignement

L'alignement des réalisations imitées par rapport au modèle est réalisé selon une procédure classique de déformation temporelle dynamique (algorithme DTW, Dynamic Time Warping). Son efficacité résulte du choix des fonctions de comparaison qui détermine un appariement optimal entre deux séries temporelles constituées par les énoncés à aligner. Ces séries temporelles sont déformées par transformation non-linéaire de la variable temporelle pour déterminer une mesure de leur similarité. Ce processus s'applique donc parfaitement au problème de la comparaison entre deux prononciations d'un même énoncé, modèle et imitation.

L'implémentation s'opère selon les étapes suivantes :

a. Un spectrogramme FFT à large bande est stocké dans une matrice temps-fréquence, dont les dimensions typiques sont 512 (nombre de créneaux de fréquences) par 2000 (nombre de valeurs de temps correspondant à un nombre standard de pixels horizontaux de l'écran).

b. Parallèlement, un spectrogramme LPC est stocké dans la matrice temps-fréquence. L'avantage de LPC par rapport à la FFT est lié à la possibilité d'utiliser une fenêtre de signal plus petite, et donc obtenir une meilleure résolution temporelle dans la comparaison des spectres relatifs au modèle et à l'imitation.

c. La matrice spectrographique (FFT ou LPC) est divisée en canaux de fréquence, de 500 Hz largeur de bande. La gamme de fréquence retenue est de 200 Hz (pour éviter le bruit de fond de l'enregistrement) à 4500 Hz (fréquence spectrale maximale utile pour la parole, à l'exclusion de la [s]). La répartition des bandes de fréquences est donc linéaire et non pas logarithmique (cette dernière, pourtant standard, s'étant avérée moins performante dans cette implémentation.)

d. Intensité : des courbes d'intensité sont calculées et stockées pour chaque canal.

Les comparaisons de spectre pour chaque déplacement sur la matrice fréquence-temps se font par rapport a ces valeurs d'intensité déjà calculées, permettant un calcul rapide.

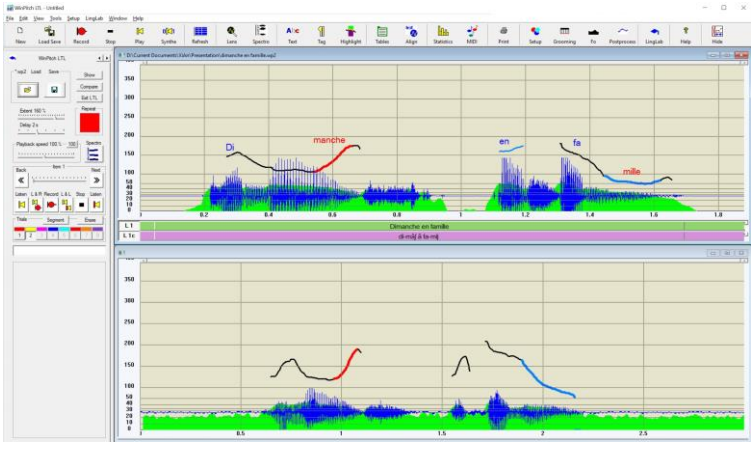


FIGURE 2 : Un exemple d'alignement automatique de segments censés être pertinents et surlignés sur la courbe mélodique modèle sur la courbe de l'apprenant. En rouge le contour de continuation C1, en bleu le contour conclusif terminal C0.

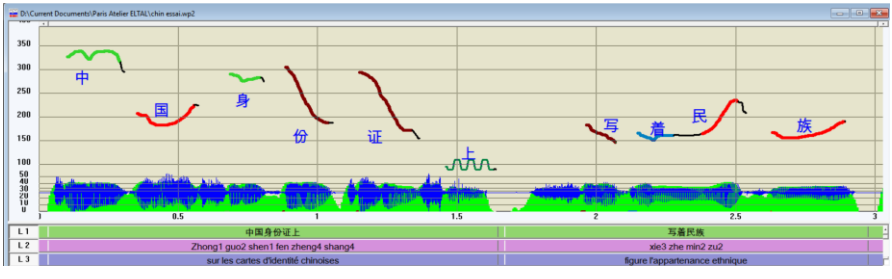


FIGURE 3 : Un exemple de surlignage couleur selon le ton d'une phrase en mandarin (Ton 1 vert, Ton 2 rouge, Ton 3 bleu, Ton 4 marron).

5. Ralenti

WinPitch LTL dispose également d'une fonction de ralenti. Basée sur la méthode TD-PSOLA (Moulines & Charpentier, 1990) et la détection des marqueurs de pitch par la méthode du peigne spectral (Martin, 1981), elle opère dans une gamme de 15% à 100% ce qui permet à l'apprenant de corréler visuellement la déplacement d'un curseur le long de la courbe mélodique avec le son de parole correspondant. En pratique, cette fonction s'avère très importante pour sensibiliser les utilisateurs non musiciens aux mouvements mélodiques, dont ils sont en général peu conscients. Le simple fait de pouvoir corréler le déplacement d'un curseur le long d'une courbe mélodique avec le segment de parole reproduit au ralenti sans déformation spectrale s'avère tout à fait convaincant pour mieux appréhender les variations mélodiques.

6. Morphing prosodique

Une autre fonction du logiciel WinPitch LTL permet de réaliser un morphing prosodique avec segmentation automatique intégrée, de manière à permettre le morphing des durées (rythmique) en plus du morphing mélodique. Cette fonction opère par segmentation automatique alignée sur les transitions spectrales (Martin et Delais, 2016) ou constante (une balise toutes les 100 ms par exemple) de manière à réaliser le morphing des durées pseudo-segmentales. Les variations mélodiques sont ensuite appliquées sur les segments correspondant du modèle à l'imitation de l'apprenant. Cette fonction permet de sensibiliser l'apprenant sur une réalisation proche du modèle utilisant sa propre voix, rendant psychologiquement ce but plus accessible.

7. Interface utilisateur

L'interface utilisateur dans sa version pour apprenant est représenté Fig. 4. Les différentes commandes et fonctions sont :

a. Lecture d'un fichier leçon, comprenant les énoncés proposés à l'apprenant.
b. Sauvegarde du fichier leçon, accompagné des différentes réalisations de l'apprenant (avec un maximum de 8 imitations par énoncé proposé).
c. Un réglage de l'échelle temporelle pour la visualisation de la courbe mélodique en temps réel.
d. Un réglage de la vitesse de playback du modèle et de l'imitation de l'apprenant.
e. Affichage d'une notice explicative en format HTML relative au modèle proposé.
f. Alignement et comparaison des segments mélodiques de l'apprenant par rapport au modèle.
g. Sélection du délai donné à l'apprenant pour démarrer son imitation (avec affichage du délai par couleurs rouge, orange et verte…).
h. Navigation dans le fichier leçon (sélection de l'énoncé précédent, suivant, ou par numéro d'ordre).
i. Sélection et affichage des réalisations précédentes de l'apprenant (maximum de 8).
j. Morphing prosodique automatique du modèle sur la réalisation de l'apprenant.

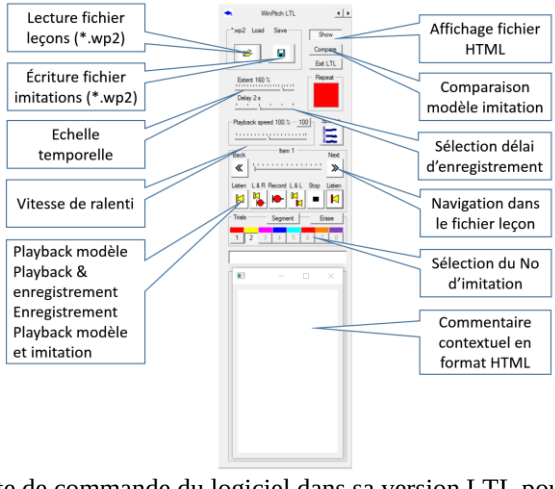


FIGURE 4 : Boîte de commande du logiciel dans sa version LTL pour apprenants

Le fichier de sauvegarde comprend les fichiers modèles accompagnés des différentes réalisations de l'apprenant, ce qui permet une évaluation et l'insertion de commentaires par l'instructeur, éventuellement à distance.

8. Conclusions

WinPitch LTL est un système ouvert et performant, bénéficiant de nombreuses fonctions auteurs pour l'élaboration de leçons d'exemples. Ces fonctions rendent l'adaptation au système de matériel déjà disponible particulièrement aisée. Ainsi des adaptations partielles de leçons de la méthode ASSIMIL ont été réalisées avec un minimum d'efforts pour l'anglais et le mandarin. Appliqué au mandarin, le logiciel utilise un codage couleur pour chacun des tons, ce qui s'avère très concluant pour l'apprentissage de cette langue. En particulier lorsque ce codage permet de se rendre compte que les tons théoriques tels que présentés dans des manuels ne sont pas toujours réalisés de la manière attendue, du fait de nombreux phénomènes de coarticulation et de sandhi.

Les points forts :

a. Système auteur ouvert, permettant l'élaboration d'exemples annotés indépendamment d'une théorie prosodique particulière.
b. Alignement automatique des segments de l'énoncé modèle sur celui de l'apprenant, avec surlignage et codage couleur des mouvements mélodiques importants.
c. Fonction de ralenti, assurant à l'apprenant une meilleure perception des mouvements mélodiques et leur relation avec la représentation graphique en courbes mélodiques.
d. Morphing prosodique, donnant à l'apprenant une réalisation selon le modèle mais utilisant sa propre voix.

Une version prochaine du logiciel devrait de plus donner un message texte des erreurs et différences des réalisations de l'apprenant, ainsi qu'une brève explication théorique.

Références

ABBERTON, E. and FOURCIN, A. (1975). Visual feedback and the acquisition of intonation. *Foundations of Language Development, eds. E.H. Lenneberg and E. Lenneberg*. 157-165. New York: Academic Press, 1975.

BENALI I., CHIARELLI R., YOO H-Y., MARTIN Ph. (2006). Enseigner la prosodie avec le logiciel WinPitch LTL : possibilités technologiques et pédagogiques. N*ouvelles technologies et éducation en milieux formel et informel*, Casablanca, 2006.

BOT, K.D. (1980). The role of feedback and feedforward in the teaching of pronunciation − an overview. *System 8*: 35-47. 1983. "Visual feedback of intonation I: effectiveness and induced practice behavior." *Language and Speech 26 (4)*: 331-350, 1980.

CHAMPAGNE-MUZARD C., BOURDAGES J. S., (1998-1993). Le Point sur la phonétique. *CLE International*, 1998/1993.

GUIMBRETIERE E. (2000). Apprendre, Enseigner, Acquérir : La prosodie au cœur du débat. *Publications de l'université de Rouen*, 2000.

JAMES, E. (1976). The acquisition of prosodic features of speech using a speech visualizer. *IRAL 14 (3)*:227-243, 1979. "Intonation through visualization." *Current Issues in the Phonetic Sciences, eds. H. A. P. Holien*. 295-301. Amsterdam Studies in the Theory and History of Linguistic Science, IV, Amsterdam: John Benjamins.

LÉON, P.R. and MARTIN, Ph. (1972). Applied Linguistics and the Teaching of Intonation. *Modern Language Journal 56 (3)*: 139-144, 1972.

GERMAIN, A. et MARTIN, Ph. (2000). Présentation d'un logiciel de visualisation pour l'apprentissage de l'oral en langue seconde. *www.alsic.org, 3, No 1*, 61-76, 2000.

JUN S-A. & FOUGERON, C. (2002). The Realizations of the Accentual Phrase in French Intonation, *Probus* 14, 147-172.

MARTIN, Ph. (1975). Analyse phonologique de la phrase française. *Linguistics*, (146) 35-68, Fév. 1975.

MARTIN, Ph. (1981). *Extraction de la fréquence fondamentale par intercorrélation avec une fonction peigne*. XIIème Journées d'Etude sur la Parole, Montréal, 223-232, mai 1981

MARTIN, Ph. (1982). Utilisation d'un visualiseur de mélodie en vue d'une didactique. *Options nouvelles en didactique du français langue étrangère.* Paris : Didier, pp. 181 – 186, 1982.

MARTIN, Ph. (2009). *Intonation du français*. Paris : A. Colin, 2009.

MARTIN, Ph. et DELAIS-ROUSSARIE, E. (2016) Comparison of three automatic speech segmentation systems, Proc. Interspeech 2016. San Francisco (submitted).

MOULINES, E. & CHARPENTIER, F. (1990). Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones. Speech Communication, vol. 9, nrs. 5/6, pp. 453-467.

STEVENS, V., SPURLING, S., LORITZ, D., KENNER, K., ESLING, J. and BRENNAN, M. (1986). New Ideas in Software Development for Linguistics and Language Learning. *CALICO Journal 4 (1)*: 15-26, 1986.

VARDANIAN, R. (1964). Teaching English intonation through oscilloscope displays. *Language Learning 14*: 109-118, 1964.

WELTENS, B. and BOT, K.D. (1984). Visual feedback of intonation II: Feedback delay and quality of feedback. *Language and Speech 27 (1)*: 79-88, 1984.

WINPITCH LTL. (2015). http://www.winpitch.com/

Vers une indexation adaptée des ressources pédagogiques sur une plateforme dédiée à l'enseignement de la Langue des Signes Française

Lucie Metz[1], Virginie Zampa[1], Saskia Mugnier[1]

(1) Laboratoire LiDiLEM, Bâtiment Stendhal, CS40700, 38058 Grenoble Cedex 9, France
lucie.metz@univ-grenoble-alpes.fr, virginie.zampa@univ-grenoble-alpes.fr, saskia.mugnier@univ-grenoble-alpes.fr

RESUME

L'enseignement de la Langue des Signes Française (désormais LSF) à distance soulève de nombreuses questions. S'il fait désormais partie intégrante du paysage numérique (projets Live-Sign, etc.[1]), nous verrons que les enseignants restent peu dotés en banque de données pédagogiques (manuels éducatifs, supports de cours, etc.). C'est pourquoi nous avons créé une plateforme en ligne permettant les échanges (supports et discussions) au sein d'une « Salle des professeurs ».
Toutefois, l'indexation des échanges (supports comme discussions) soulève de nombreux problèmes compte tenu des différents déterminants sociétaux que nous exposerons rapidement. Nous verrons que malgré leur grand degré de précision, les modèles d'indexation existants ne sont pas utilisables dans notre situation. Nous présenterons donc nos solutions pour rendre accessible l'indexation (en limitant le recours à l'écrit) de supports pédagogiques (souvent au format vidéo) pour notre projet de plateforme pour l'enseignement de la LSF.

ABSTRACT

For an appropriate indexation of pedagogical resources for French Sign Language Teaching platform web-design
Distance education for French Sign Language (LSF) raises many questions. Nowadays, this LSF Teaching forms an integral part of the French digital landscape (Live-Sign project, etc.[1]). Nevertheless, we will see that teachers have little access to pedagogical data banks (pedagogical handbooks, teaching materials, etc.). For this reason, we created a web-design teaching platform allowing educational exchanges (teaching materials and discussions) inside a "Teacher's room".
However, the indexation of these exchanges shows many problems taking into account the various societal factors, which we will describe quickly. We'll see that existing indexation models aren't applicable in our case, despite their high level of precision. Consequently, we'll expose our indexing and information search solutions for our platform with the aim of making information more accessible to teachers.

MOTS-CLES : Ressources pédagogiques, Indexation, LSF, Enseignement à distance

KEYWORDS: Pedagogical Resources, Indexation, French Sign Language, Distance Teaching

[1] LiveSign : http://live-sign.com/ dernière consultation le 9 janvier 2016 ; Lingueo : http://www.lingueo.fr/nos-formations/cours-langue-des-signes-francaise.html dernière consultation le 9 janvier 2016 ; Sourdoués : http://www.signesetformations.com/formation-a-distance-pour-les-sourds/ dernière consultation le 9 janvier 2016.

LUCIE METZ, VIRGINIE ZAMPA, SASKIA MUGNIER

1 Spécificités de la LSF et contexte d'enseignement

1.1 Spécificités de la LSF

La LSF demeure à ce jour une langue sur laquelle il existe bon nombre d'idées reçues. Cette partie vise à les dissiper et à ancrer notre projet dans la réalité de notre terrain. Ainsi, nous verrons quelles spécificités rendent cette langue singulière ; et, quelles particularités font des enseignants de LSF un public aux contextes d'enseignement variables.

La LSF a obtenu son statut officiel de langue à part entière récemment puisqu'il date de la Loi 2005-102 du 11 février 2005 (article L.312-9-1) pour l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées. Ce statut officiel fournit un cadre institutionnel censé favoriser son enseignement et son apprentissage. Mais il faut savoir que les Langues des Signes sont des langues à tradition orale, ne disposant pas de forme écrite institutionnelle. Le français écrit n'est donc pas sa forme écrite. Ceci induit dans le cas des Environnements Informatiques pour les Apprentissages Humains (désormais EIAH) une adaptation qui serait valable pour les langues orales. Effectivement, dans la plupart des outils existants l'oralité n'est pour ainsi dire pas prise en compte ; or, dans notre situation le traitement de l'oralité est non seulement nécessaire mais primordial ; et ne peut se contenter d'une « simple » traduction des informations textuelles généralement disponibles en vidéos signées. De plus, les langues des signes sont des langues qualifiées de visuo-gestuelles de par les canaux de communication qu'elles empruntent (Moody et al., 1983) et leurs locuteurs s'expriment dans un espace de signation tridimensionnel (Millet, 1997). Ces particularités font de la LSF une langue singulière dans le cadre des EIAH.

1.2 Particularités du public visé : les enseignants de LSF

Contrairement aux autres langues pour lesquelles la plupart des enseignants possèdent le CAPES ou l'agrégation, le statut des enseignants de LSF dans la société est assez variable. Golaszewski (2010) souligne qu'en 2010, la majeure partie du vivier existant pour le 2nd degré provenait du réseau associatif ou des personnels de santé. Ceci s'explique par la mise en place récente du CAPES de LSF en 2009[2] et ne reflète donc pas la situation professionnelle de la majorité des enseignants de LSF exerçant à ce jour. Elle précise également qu'en 2007 près de 200 enseignants du primaire ont été recrutés soit par concours soit sur contrat suite aux décisions politiques de diffusion de la LSF. La création du CAPES et celle du Certificat Complémentaire « Enseignement en LSF » (« Certification Complémentaire », 2009) ont permis, toujours selon Golaszewski (2010), de consolider les ressources humaines existantes. La mise en place de ces certifications a en revanche renforcé les inégalités de statut entre les enseignants. En effet, une grande partie des enseignants de LSF demeure dans une situation professionnelle précaire (vacations, multiplicité de Contrats à Durée Déterminée) quand une petite partie est en situation plus stable (titulaire de l'Éducation Nationale ou Contrats à Durée Indéterminée). À cela s'ajoute que ces enseignants sont répartis inégalement sur le

[2] CAPES LSF : http://www.cndp.fr/ressources-lsf/capes/pdf/joe_20090627_0147_0057.pdf dernière consultation le 29 janvier 2016

territoire et qu'ils demeurent de fait isolés géographiquement et pédagogiquement les uns des autres. L'hétérogénéité du public se ressent également dans le rapport particulier qu'une partie des enseignants sourds de LSF entretient avec l'écrit. En effet, il est reconnu que certains ont une maîtrise variée et variable du français écrit. Ce dernier point a notamment été confirmé par les enseignants avec lesquels nous avons échangés lors de nos recueils de données.

Le public que nous visons est donc assez hétérogène et l'ensemble des particularités présentées fait de la LSF une langue singulière dans le cadre des EIAH. L'ensemble de ces spécificités nous amène à nous interroger sur les solutions possibles à l'indexation de supports pédagogiques pour l'enseignement de la LSF. Pour répondre à cette question, nous montrerons que, compte tenu des spécificités du public, il n'est pas possible de suivre les modèles d'indexation existants. Nous proposerons des solutions en accord avec les besoins des acteurs de terrain, à savoir les enseignants de LSF. Pour ce faire, nous avons décidé de proposer une plateforme en ligne correspondant le mieux possible aux besoins de ces enseignants tant d'un point de vue « social » que pédagogique.

2 Point de vue théorique

2.1 Qu'indexons-nous ?

Avant tout il est primordial de définir la notion d'indexation pédagogique. Pour cela, nous nous rapprochons de la définition de Loiseau (2009 : 16) pour qui une « *indexation* [est] *effectuée selon un langage documentaire permettant à l'utilisateur de rechercher des objets pour les utiliser dans un enseignement* ». En effet, la plupart du temps la problématique est la suivante : l'enseignant ajoute un texte dans le système. Mais ce qui l'intéresse, c'est avant tout de pouvoir trouver un texte (en fonction des notions qu'il souhaite aborder, du public visé par son enseignement, du niveau du public, etc.) et donc de chercher efficacement celui qui répondra à ses attentes (Figure 1).

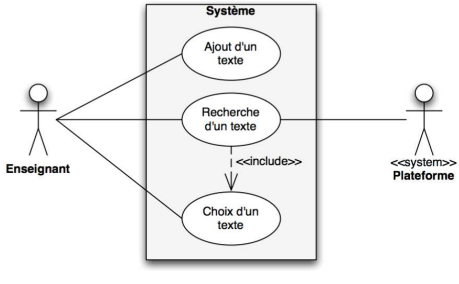


FIGURE 1 Image extraite de Loiseau, (2009 : 8)

Or, dans notre cas la recherche d'objets soulève quelques questions. En effet, les enseignants n'ajoutent que rarement du texte et ne recherchent qu'encore plus occasionnellement des documents textuels. Le langage documentaire à adopter est également délicat. Comme nous l'avons évoqué, les enseignants n'ont qu'une maîtrise partielle de l'écrit ; il est alors judicieux de s'interroger sur les possibilités existantes pour fournir une recherche efficace des supports pédagogiques. Bien souvent les enseignants sont amenés à déposer des supports de type vidéo, image, Présentation Assistée par Ordinateur, voire dans de rares cas des supports textuels.

2.2 Les différents systèmes d'indexation existants

En 2009, Loiseau a dressé un tour d'horizon assez complet des différents modèles d'indexation de ressources pédagogiques existants. La dynamique de standardisation qui était en cours cherchait à uniformiser les pratiques d'indexation en proposant un modèle qui pourrait répondre de façon précise à tous les types de support. Nous nous focaliserons seulement sur quelques modèles qui se caractérisent soit pour leur fréquence d'utilisation ou leur institutionnalisation. Le *Sharable Content Object Reference Model* (SCORM) a connu son avènement au début des années 2000 quand un certain nombre de Learning Management Systems (désormais LMS) ont décidé de l'appliquer comme standard d'indexation. Les LMS sont des outils-auteurs permettant de créer et d'organiser des contenus pédagogiques. C'est notamment le cas des environnements numériques d'apprentissage tels Moodle, Chamilo, Dokeos, Claroline Connect, etc.

Ainsi, le SCORM est certainement un des plus utilisés dans le cadre des LMS. Toutefois, il est davantage utilisé pour le suivi des apprenants plutôt que dans l'optique d'une recherche efficace pour l'échange de supports pédagogiques entre enseignants. Qui plus est, ce standard n'est plus mis à jour depuis 2009 et est progressivement remplacé depuis 2013 par le standard Tin Can API (ou xAPI) (Fraysse, 2015). Le SCORM n'est donc pas pertinent dans notre cadre.

Un des plus célèbres modèles reste le *Learning Object Metadata* (ou LOM), un standard américain dont la dernière mise à jour date de 2002, qualifié de « *modèle central dans le panorama des métadonnées pédagogiques* » (Loiseau, 2009 : 105). Si ce standard est reconnu pour être très complet et favorisant l'interopérabilité des ressources (Gómez de Regil, 2004), il reste inaccessible pour un public de non-initiés (Figure 2).

Ce standard a été repris en vue d'une normalisation adaptée au modèle francophone et a vu naître le profil d'application LOMFR[3] dès 2006 (Éducation Nationale, 2006). Ce dernier se fonde sur le standard LOM mais par souci de compatibilité est aussi conforme à d'autres standards tel *Dublin Core Element Set* (Loiseau, 2009 : 95). C'est un des premiers modèles à avoir vu le jour ; il se compose d'un ensemble de 15 descripteurs de base pour les ressources numériques. Il est généralement décrié pour son manque de précision (Gómez de Regil, 2004). Dans notre cas, il ne correspond pas aux types d'informations pertinentes dont nous avons besoin. Les profils d'application SupLOMFR et ScoLOMFR s'inscrivent dans la lignée du LOMFR.

[3] Carte heuristique du LOMFR v1.3 de 2014 :http://www.lom-fr.fr/scolomfr/fileadmin/user_upload/docs_utiles/ScoLOMFR-2014-11-Freeplane.pdf

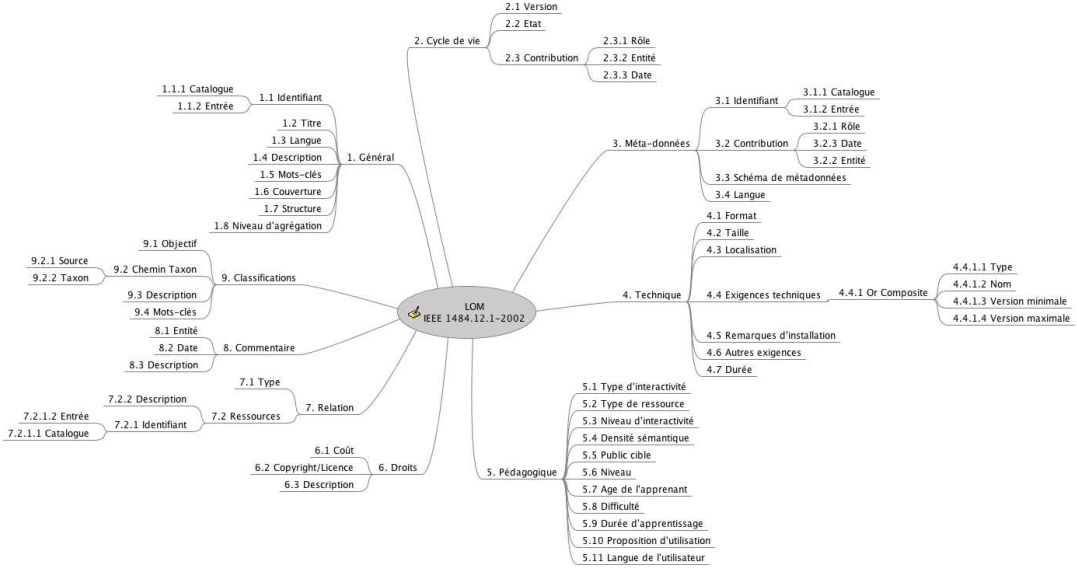


FIGURE 2 Aperçu général du LOM (Gómez de Regil, 2004)

Compte tenu des exigences de terrain, les enseignants devant remplir eux-mêmes les métadonnées sur les supports qu'ils souhaitent déposer, ces normes et standards ne peuvent être utilisés tels qu'ils sont présentés ; soit parce qu'ils ne correspondent pas aux attentes soit trop complexes à compléter. De la même façon, il ne serait pas judicieux d'extraire une partie d'un standard puisqu'il perdrait là tout son intérêt d'interopérabilité.

Si ces modèles ne correspondent ni à notre public ni à notre objet, nous souhaitons malgré tout fournir une banque de données de ressources pédagogiques numériques cohérente, utilisable et de bonne qualité.

2.3 Langue des Signes Française et indexation, quelles ambitions ?

Les chercheurs s'accordent sur la révolution des pratiques induite par l'accès massif aux nouvelles technologies (caméra, ordinateurs, smartphones, etc.) qui permettent la diffusion de contenus directement en LSF notamment grâce à la vidéo. Golaszewski, (2010) présentait en 2010 les apports de la « LSF numérique » notamment par rapport à la vidéo qui selon elle « *révolutionne la pédagogie et la didactique de la LSF* » en faisant le parallèle avec l'écrit des langues vocales (p.65) ; de même, Dalle, (2011) explique que la vidéo « *va remplir en partie le rôle d'une écriture, pour les fonctions de trace et de mémorisation* » (p.92).

Toutefois, il se pose la question de l'indexation de ces vidéos. En effet, il n'est pas à ce jour possible de pouvoir faire une recherche au sein d'une vidéo, tout comme il n'est pas encore possible de rechercher une information directement en LSF via un système de reconnaissance automatique des signes (Dalle, 2011).

La récence des technologies ajoutée au statut récent de la LSF explique le décalage qui subsiste entre reconnaissance automatique de la parole et la reconnaissance automatique des signes (Braffort & Dalle, 2012). Les projets existants restent donc bien souvent au stade de prototypes expérimentaux non-utilisables à plus grande échelle. Braffort & Dalle regrettent que malgré une utilité certaine, « *il n'existe actuellement que peu de solutions opérationnelles sur l'accès à des contenus en LS (requête directement en LS, indexation, navigation en LS)* » (Braffort & Dalle, 2012 : 287).

Ainsi, les apports potentiels des nouvelles technologies dans l'enseignement de la LSF se distinguent sous plusieurs aspects (visualité, interactivité, etc.). Comme nous avons pu le noter, les modèles existants ne nous permettent pas, d'une part d'indexer les ressources pédagogiques au format vidéo et, d'autre part, de les chercher librement et directement en LSF.

3 Les pratiques déclarées des enseignants

Si les impacts potentiels des EIAH dans l'enseignement de la LSF sont indéniables en théorie, nous avons cherché à savoir ce qu'il en était pour les acteurs de terrain. Pour ce faire, nous avons interrogé dix-huit enseignants avec un questionnaire en ligne dont une partie a ensuite accepté de participer à un entretien. Après avoir présenté brièvement les méthodes utilisées lors des recueils de données, nous exposerons les pratiques numériques déclarées par les enseignants interrogés et leurs attentes et conditions d'utilisation émises quant à la création d'une plateforme en ligne.

3.1 Méthodologie des recueils de données

Le questionnaire en ligne a été diffusé via la plateforme Moodle du 25 juillet 2015 au 18 décembre 2015. Les prises de contact se sont faites par mail via des listes de diffusion visant la communauté Sourde[4] et des communautés d'enseignants (associations). Suite à cette diffusion vingt-cinq personnes nous ont contactées pour participer. Trois n'ont pas répondu à nos relances et n'ont jamais donné suite à leur prise de contact. Quant aux vingt-deux personnes ayant tenté de se connecter, dix-huit ont finalisé leur participation. Quatre personnes ont donc décidé d'abandonner dont deux qui nous ont fait part de leurs raisons. Ces désistements s'expliquent notamment par la complexité d'accès au questionnaire (nécessité de se faire connaître et de se connecter à la plateforme Moodle pour y accéder notamment), et le manque de temps pour y répondre.

[4] L'utilisation de la majuscule à « Sourd » est une convention permettant de désigner la communauté signante des sourds (Millet, 2003).

L'objectif de cette enquête était de pouvoir : cerner l'utilisation qu'ils font des outils numériques dans leurs pratiques pédagogiques (en préparation, pendant et/ou après les cours) ; connaître leurs besoins « technologiques/numériques » dans le cadre de leurs pratiques ; mais aussi de trouver un vivier d'enseignants susceptibles de participer durablement à la plateforme.

Les entretiens quant à eux se sont déroulés du 2 décembre 2015 au 11 février 2016. Pour ce faire, nous avons décidé de recontacter les participants au questionnaire en ligne afin de savoir lesquels souhaitaient poursuivre leur investissement dans notre projet. Les rencontres se sont faites en présentiel ou via Skype et dans les deux cas avec ou sans interprète LSF-français pour les enseignants sourds. Dix enseignants ont été rencontrés individuellement pendant une trentaine de minutes.

L'objectif de ces rencontres était d'une part de pouvoir affiner les réponses des enseignants formulées lors du questionnaire et, d'autre part, de mieux connaître leurs impressions sur les outils numériques appliqués aux Langues des Signes. La question qui nous occupait étant : quelles difficultés rencontrent les enseignants de LSF lors de l'utilisation d'outils numériques ? Les réponses des enseignants sont en cours de traitement. Cela dit, il est tout de même possible de noter plusieurs points de convergence entre les différents participants.

3.2 Pratiques actuelles

Les participants au recueil nous ont paradoxalement fait part de leur grand intérêt et d'un recours régulier aux outils numériques mais aussi de leur scepticisme sur *l'utilisabilité* de ces outils. Il est important de souligner qu'une très large majorité d'entre eux utilise l'ordinateur que ce soit en préparation, pendant ou en correction de leurs cours. La plupart ne se considère pas comme technophile mais ils reconnaissent la dimension indispensable de l'ordinateur dans leurs enseignements. C'est pourquoi ils expliquent l'utiliser très régulièrement sans pour autant se sentir particulièrement à l'aise. Lors des entretiens, quand nous lui avons demandé si elle était à l'aise avec l'informatique et les nouvelles technologies une des enseignantes nous a donné la réponse suivante :

> « oui bah au début franchement non c'était pas le cas mais j'ai suivi des cours j'ai été formée euh on m'a aidée euh et puis j'ai pratiqué *[…]* euh je pense que quand même la condition c'est d'aimer un petit peu l'informatique hein faut quand même bien aimer ça euh moi franchement je suis pas f- je suis pas une férue de l'informatique mais j'en ai réellement besoin c'est des outils extrêmement utiles pour la pédagogie *[…]* »

Les logiciels les plus fréquemment utilisés par les enseignants interrogés sont des logiciels de présentation, de montage/lecture vidéo, et traitement de texte et d'image. Tous les enseignants déclarent utiliser des supports de type vidéo et image. D'autres supports : texte, jeux, dictionnaire, etc. sont également mentionnés. Il faut donc tenir compte de ces deux points (typologie des supports pédagogiques et logiciels utilisés) dans notre outil puisque les enseignants sont généralement les créateurs des supports pédagogiques qu'ils vont être amenés à déposer. Ils doivent être en mesure d'ajouter efficacement tous ces types de supports (créés comme récupérés) dans notre outil.

Le second volet des échanges qui nous intéressait concerne la possibilité d'échanger directement avec ses pairs sur des problématiques liées à l'enseignement ainsi que la mise en place d'un cadre de mutualisation des réflexions pédagogiques en LSF. Nous souhaitions connaître la fréquence à laquelle les enseignants échangent avec leurs pairs (si c'est le cas bien entendu) ; les modalités de ces échanges lorsqu'ils se produisent et si ces échanges sont suffisants pour eux. La moitié d'entre eux déclare échanger rarement (moins d'une fois par mois à jamais). Ces échanges se font principalement « peau » (en présentiel) voire sur Skype mais les échanges asynchrones existants (mail, forums, réseaux sociaux, etc.) se prêtent généralement assez mal à une modalité visuo-gestuelle. Nous leur avons également demandé dans quelle mesure cette fréquence leur suffisait ; leurs réponses montrent que ces échanges ne les satisfont pas mais ils expliquent également qu'ils ne peuvent pas rencontrer en présentiel plus de personnes faute de temps. Il est donc primordial de proposer un outil de mise en commun des réflexions directement en vidéo et proposant une visualisation directe depuis la plateforme.

3.3 Attentes et conditions d'utilisation

Si les enseignants ont très majoritairement (16 sur 18) sollicité la mise en place d'un outil numérique pour l'enseignement de la LSF, ils ont toutefois exprimé leurs attentes en la matière et plusieurs conditions. Selon eux, leurs besoins se situent principalement au niveau des échanges avec leurs pairs ; que ce soit de l'échange/renouvellement de supports pédagogiques comme la possibilité d'interagir avec ses pairs en vue d'une réflexion commune sur leurs pratiques pédagogiques. Ils attendent de ce type d'outil un gain de temps en préparation et en recherche de support et un enrichissement de leurs contenus pédagogiques.

Par ailleurs, selon eux, l'outil devra limiter le recours à l'écrit et être facile d'utilisation. Ils évoquent les problèmes récurrents qu'ils ont pour échanger des supports vidéo (difficultés techniques). Ils s'interrogent également sur la non-anonymisation des vidéos et le possible jugement négatif de leurs supports par leurs pairs ; et enfin sur la réticence de leurs pairs à partager les supports en leur possession. En effet, cette dimension a été soulevée tant dans le questionnaire que lors des entretiens où chacun a exprimé sa crainte de donner sans avoir de supports en retour. Cette dernière appréhension est d'autant plus importante dans notre situation que les enseignants de LSF sont connus pour être assez peu habitués aux normes de citation des sources et craignent souvent que d'autres enseignants s'approprient indûment leurs supports.

4 Prototype expérimental et solutions envisagées

Les difficultés techniques soulevées par les acteurs de terrain ont confirmé les doutes que nous avions sur le recours à un outil existant de type LMS (Moodle, Claroline Connect, Dokeos, Wiki, etc.). En effet, les systèmes ne sont qu'assez peu intuitifs ; d'une structure souvent jugée complexe (beaucoup de « clics » sont nécessaires pour des actions considérées comme simples) ; et à laquelle il faut ajouter une hyper présence de l'écrit. Toutes ces dimensions en font des outils non utilisables en l'état pour nos publics.

Nous avons donc décidé de proposer un outil qui correspondrait mieux à leurs attentes et où le recours à l'écrit serait minimaliste. L'espace réservé aux enseignants dans notre plateforme s'intitule « Salle des Professeurs ». Cette salle se divise à son tour en trois espaces : la boîte à outils pratique, la boîte à outils théorique et un dernier espace de mise en commun des réflexions.

4.1 Ergonomie

Pour ce prototype nous avons choisi de proposer une interface la plus épurée possible. Lorsque l'enseignant se trouve sur la page d'accueil de la Salle des Professeurs, les informations sont présentées selon trois modalités (Figure 3) : l'intitulé de l'espace est en français écrit, il y a un descriptif des possibilités offertes par l'espace en LSF sous forme vidéo et le bouton pour accéder à l'espace est précédé d'une icône qui lui est propre : une caméra pour l'espace de mise en commun des réflexions, un signe plus pour accéder au dépôt de supports dans la boîte à outils pratique, etc. chaque icône est explicitée dans la description en LSF qui lui correspond.

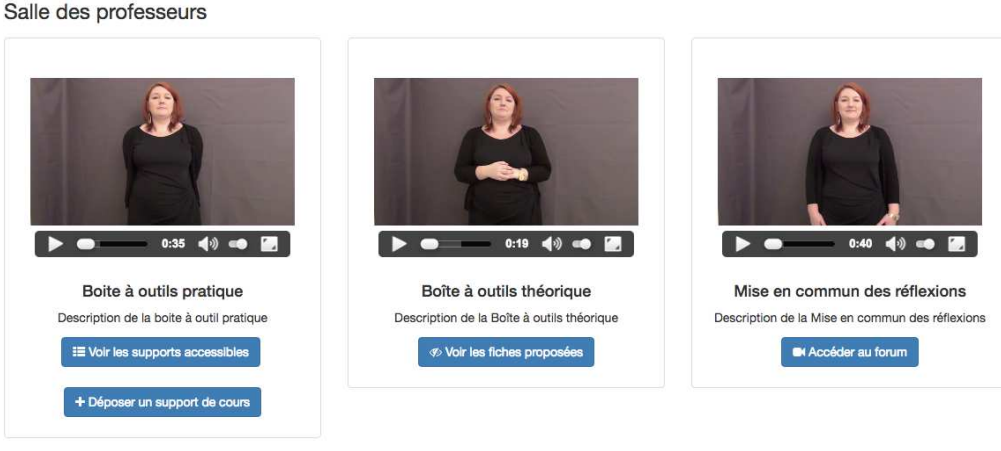


FIGURE 3 : Copie d'écran de la page d'accueil du prototype de la plateforme

Le fonctionnement est le même quelles que soient les pages de la plateforme. Pour chaque action possible l'enseignant dispose à la fois de l'information textuelle et de l'information iconique (à gauche de l'information textuelle sur chacun des boutons cliquables) de même qu'une vidéo-tutoriel expliquant les différentes possibilités offertes dans l'espace affiché. Nous nous focalisons ici plus particulièrement sur la boîte à outils pratique et l'espace de mise en commun des réflexions.

4.2 Boîte à outils pratique et indexation des supports pédagogiques

Concernant les choix que nous avons faits pour l'indexation des supports pédagogiques, comme nous l'avons vu, il n'est pas possible de contraindre les enseignants à utiliser des modèles d'indexation existants compte tenu de leur complexité et de leur incompatibilité en l'état. Toutefois, il nous fallait trouver une solution pour permettre une recherche efficace des contenus. Nous avons élaboré pour ce premier prototype un système d'étiquettes.

4.2.1 Des étiquettes, pourquoi ?

Nous avons fait le choix de proposer dans notre prototype expérimental un système d'étiquettes catégorisées (voir la Figure 4). Nous proposons des étiquettes car elles permettent aux enseignants de « réfléchir en mots-clés » ce qui limite le recours aux morphèmes grammaticaux (préposition, déterminant, pronom, conjonction, etc.) mais aussi limite l'emploi de formes verbales conjuguées complexes. Chaque étiquette ajoutée par l'enseignant doit s'intégrer au sein d'une des catégories proposées. En effet, ces catégories sont fixes et permettent de guider les enseignants dans les informations qu'ils estiment pertinentes. Il existe cinq catégories d'étiquettes : public ; niveau ; thème ; objectif ; et autre.

Public permet de définir les populations visées par l'apprentissage. Pour la première version, nous avons décidé d'en suggérer quelques unes même si les enseignants n'y sont pas contraints et peuvent en créer de nouvelles (enfant, adolescent, adulte, langue maternelle et langue seconde).

Compte tenu des disparités terminologiques entre les organismes de formation, il était complexe de proposer des étiquettes pour le *niveau*. Nous avons décidé de laisser cette liste vierge de proposition. Les enseignants sont libres d'y ajouter les niveaux de compétence du Cadre Européen Commun de Référence pour les Langues des Signes (Conseil de l'Europe, 2002) ou la fréquente catégorisation : débutant, moyen, avancé.

Thème vise l'emploi, le contexte dans lequel le support est utilisé : couleurs, temporalité, se présenter, etc.

Objectif cible les compétences à acquérir ou en cours d'acquisition avec ce support, par exemple des libellés tels que vocabulaire, syntaxe, etc. sont envisageables.

Pour les libellés qui n'entrent dans aucune des catégories, nous proposons une catégorie *autre*.

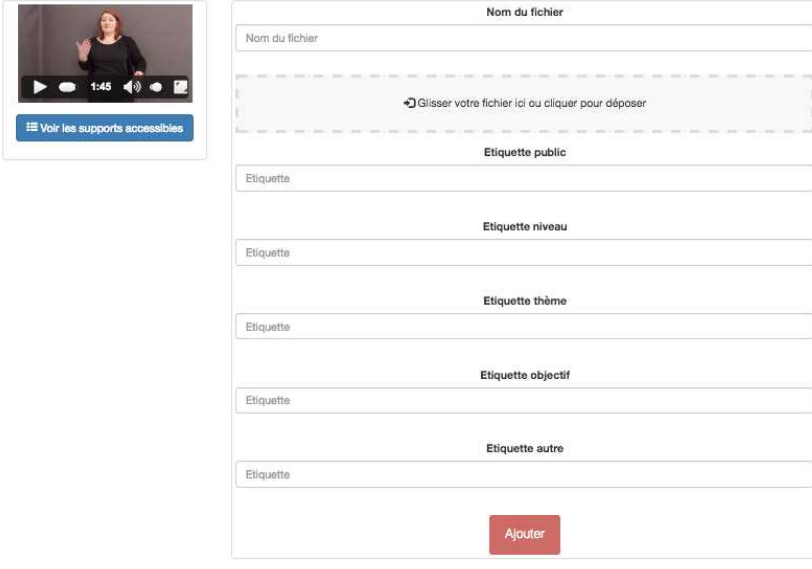


FIGURE 4 : Copie d'écran du prototype de la plateforme dans l'espace de dépôt de support

L'enseignant n'est pas limité dans le nombre d'étiquettes pour chaque catégorie. En effet, il peut tout à fait déposer un support qu'il utilise auprès d'adultes et d'adolescent (i.e. *public*) ou dans différentes classes de niveaux. De même, un support pictural pourrait tout aussi bien servir dans deux thèmes distincts. Un enseignant peut aussi estimer qu'une catégorie (ou plusieurs catégories) n'est pas pertinente pour le support et choisir de ne pas remplir toutes les catégories disponibles. Cette liberté de choix dans l'intitulé des étiquettes est volontaire, l'objectif est d'inciter les enseignants à proposer des étiquettes correspondant réellement à leurs besoins. Pour limiter les risques de doublons, dès que l'enseignant rempli un nouveau libellé, un système d'auto-complétion lui suggère tous les libellés contenant la chaîne de caractère qu'il vient de saisir.

Ce système d'étiquette doit nous permettre de mieux cerner les libellés fréquemment employés par les enseignants et qui seront donc nécessaires à la plateforme. En effet, l'objectif, pour la prochaine version (plus largement diffusée), est de mettre en place un système d'indexation différent s'appuyant sur les libellés saillants ou indexations personnelles (folksonomie) issues de ce premier prototype. Ainsi, nous nous rapprocherons le plus possible des demandes de terrain (clarté d'utilisation et recours limité aux informations textuelles).

4.2.2 Et après ?

Comme nous l'avons évoqué, notre volonté est de réduire l'emploi du texte dans le système d'indexation à la réelle utilisation que les enseignants en font. Pour cela, nous déterminerons un

système de recherche et d'indexation se rapprochant d'un système de vocabulaire contrôlé conçu à partir de la folksonomie issue de l'expérimentation en cours. Il prendra soit une forme écrite pour les notions couramment maîtrisées (pour les niveaux par exemple comme A1, A2, B1, etc.), soit en utilisant un système d'icônes de recherche ou directement en LSF en fonction du degré de complexité des concepts.

4.2.3 La recherche de contenus

Les enseignants peuvent ensuite consulter les supports déposés. L'ensemble des supports est présenté sous forme de tableau recensant les informations générales sur le support (nom du support, date et heure de dépôt, auteur), les libellés des différentes sortes d'étiquettes ayant chacune une couleur spécifique (la même que celle s'affichant dans les onglets de dépôt) et la possibilité de télécharger le support voire de le supprimer si l'enseignant en est l'auteur (Figure 5).

FIGURE 5 : Résultat d'une recherche de support sur la plateforme

Pour l'instant, la recherche se fait grâce à un champ disponible au-dessus du tableau sur les informations générales du support et chaque libellé d'étiquette. Le résultat de la recherche s'affiche en temps réel. En revanche, une fois que nous aurons stabilisé les libellés d'étiquettes la recherche pourra se faire directement à partir des icônes, en LSF voire toujours en français écrit. Les entretiens que nous mènerons à la fin de notre expérimentation nous permettront notamment de décider quel outil de recherche est le plus ergonomique.

4.3 Indexation au sein de l'espace de mise en commun des réflexions

L'espace de mise en commun des réflexions pédagogiques se constitue d'un forum vidéo. En effet, il était essentiel que les enseignants puissent interagir directement en LSF que ce soit pour l'enregistrement comme la visualisation des contributions. Concernant la recherche au sein du forum vidéo, elle se fait à l'heure actuelle en français écrit sur les noms des participants, le sujet, la description (actuellement en français écrit mais qui sera à terme directement en LSF), et, plus tard pourra aussi se faire à partir des mêmes « mots-clés » que ceux proposés par l'outil de recherche des supports pédagogiques.

5 Perspectives

L'expérimentation, débutée le 2 mars 2016 avec les enseignants ayant acceptés de s'investir dans le projet (9 sur les 18 ayant participé à la première enquête) est toujours en cours à l'heure actuelle. Elle devrait nous permettre de mieux cibler les adaptations nécessaires. Cette expérimentation s'inscrit réellement dans une démarche de co-construction en accord avec les acteurs de terrain en vue d'une diffusion plus massive de la plateforme.

Références

BRAFFORT, A., & DALLE, P. (2012). Accessibilité du cyberespace : Langue des Signes. In *NET.LANG Réussir le cyberespace multilingue* (C&F éditions, Vol. NET.LANG réussir le cyberspace, p. 268-289). Mayaa. Consulté à l'adresse http://net-lang.net//externDisplayer/displayExtern/_path_/netlang_FR_pdfedition.pdf

CERTIFICATION COMPLEMENTAIRE. (2009). Consulté le 24 février 2015, à l'adresse http://www.education.gouv.fr/cid50068/menh0929050n.html

CONSEIL DE L'EUROPE (2002). Cadre Européen Commun de référence pour les langues. Strasbourg : coll Didier. Dernière consultation le 16 février 2016 http://alfeda.net/pdf/langue%20des%20signes.pdf

DALLE, P. (2011). Technologies de l'information et de la communication au service de la LSF. In *Contacts sourds-entendants N° 6 - Grandir et apprendre en langue des signes - Oui, mais comment ?* (p. 89-109). L'Harmattan.

ÉDUCATION NATIONALE. (2006). LOMFR (Learning Object Metadata) Consulté le 9 février 2016, à l'adresse http://eduscol.education.fr/numerique/dossier/archives/metadata/ressources-educatives-numeriques/lomfr-learning-object-metadata

FRAYSSE, S. (s. d.). SCORM en 10 questions. Consulté à l'adresse http://fraysse.eu/scorm-en-10-questions/

GOLASZEWSKI, M. (2010). *La langue des signes française - À l'école, au collège et au lycée*. CNDP - Chasseneuil-du-Poitou. Consulté à l'adresse https://www.reseau-canope.fr/notice/la-langue-des-signes-francaise.html

GOMEZ DE REGIL, R. M. (2004, novembre 16). Présentation des standards : (LOM) - Learning Object Metadata [text]. Consulté 25 septembre 2014, à l'adresse http://www.enssib.fr/bibliotheque-numerique/notices/1237-presentation-des-standards-lom-learning-object-metadata

LOISEAU, M. (2009, décembre). *Elaboration d'un modèle pour une base de textes indexée pédagogiquement pour l'enseignement des langues*. Université Stendhal - Grenoble 3. Consulté à l'adresse http://tel.archives-ouvertes.fr/docs/00/46/18/79/PDF/These-MathieuLoiseau.pdf

MILLET, A. (1997). Réflexions sur le statut du mouvement en LSF - aspects lexicaux et syntaxiques. *LIDIL, 15*(Langues gestuelles, quels enjeux pour les Sourds ?).

MILLET, A. (2003). « Les représentations de la LSF – Comment penser un sujet sourd bilingue et biculturel ». *Nouvelle Revue de l'AIS Langue des signes française – Enjeux culturels et didactiques*, (Ed. du Cnefei), 63-72.

MOODY, B. (1983). *La langue des signes. Tome 1: Histoire et Grammaire*. Vincennes, France: International Visual Theatre - Centre Socio-Culturel des Sourds.